STAD94: The Champions Analytics

Subagan Kamaleswaran 1002163712 / Hamish Rajiv 1003136117/ Bavalan Thangarajah 1002194564 / Karsan Uthayakumar 1002585936

December 3, 2019

Abstract

In present day, many different industries collect a vast amount of data in numerous different areas for the purpose of analyzing and learning. As data analysis became a popular tool, many sports teams began collecting data to provide themselves with insight on their performance and understand how to improve their team in order to gain an advantage against their competitors. The sports chosen were hockey and basketball, specifically analyzing the data from the National Hockey League (NHL) and National Basketball Association (NBA) respectively. Both sports have similar parameters thus making the specific models created for each comparable. Parameters were created based on research and intuition. The Data used to create the model had been collected from official websites of the NHL and NBA. Using ordered logistic regression for both sports, the most beneficial variables were chosen to create a model. In order to determine the most beneficial variables statistical modeling techniques such as backward and forward elimination along with intuition were used. After eliminating variables to create the most influential model for the NHL, we determined Elite Forward, Elite Defense and Conference Quarter-Final Win Percentage were significant. On the other hand, the most influential model for NBA included parameters such as Elite Defense, Elite Forward, and 3-point percentage.

Background and Significance

In sports, there are many different philosophies on how a team should be constructed. To answer the question of how to construct a championship-caliber team many sports teams turn to data analytics tools. Data analysis is important to sports because it helps determine patterns and gives valuable insight to help teams understand the obstacles they may face and provide a solution. For example, an NHL team can use the data collected on their teams to determine an offensive defenseman with high puck possession and fewer giveaways can benefit their team. Data collection and analyzation can also benefit a coach as they learn and improve their decision-making to coach the best play for a specific situation.

Sports analysis also plays an important role in improving a player's performance. The data collected on a specific player can give insight into determining if the player would be more productive playing with as a result that would help enhance the performance of the entire lineup. Athletes can also understand how exhaustion can affect their workout during practice or during training. Furthermore, analyzing data can also help prevent injuries to players as they become more knowledgeable on avoiding certain plays that can cause injuries. For instance, in the 2013-2014 season the NHL removed a 76 year old touch-icing system and introduced a hybrid icing system. The change prevents players who iced the puck from racing with the opposing players to touch the puck first and avoid an icing call. With hybrid icing, if a team ices the puck the play will be whistled down if the opposing player is closer to the puck. The change was made because data analytics found many players have career-ending injuries trying to race to the puck first.

Data examination can also increase a team's revenue by not only creating the best teams but also help deliver a better experience for fans. Finally, Data analysis is important for teams to know by how much they can increase tickets and merchandise sales in order to not affect attendance.

Many teams in the NHL and NBA currently use a similar system to help improve their team's chances of winning a championship. The Golden State Warriors, an NBA team that has won two Championship titles in the last three years, are known for being pioneers in using technology in basketball. The Warriors General Manager Kirk Lacob has brought analytics, machine learning, and data science into the basketball court. Kirk Lacob said "A huge part to understanding why we need analytics in sports is figuring out the right question to ask." Data is changing the way many teams play the sport because they realize it can help shape the next championship team.

This report explores two different statistical models predicting which parameters help enhance an NHL or NBA team's chances of winning a Stanley Cup or Larry O'Brien trophy respectively.

Exploratory Data Analysis

This data set contains 256 observations and 31 variables. The variables are listed below:

Independent Variables	
Rk	Rank finished the playoffs
Team	Name of the NHL team
Gp	Games played: includes games played after train deadline and all playoff games played
W	Number of games won after trade deadline + playoffs
L	Number of games lost after trade deadline + playoffs
W-L%	Number of games won after trade deadline + playoffs (W) divided by the total number of games played (GP)
G	Total number of goals scored in the total number of games played (GP)
GA	Total number of goals scored against in a total number of games played (GP)
DIFF	Total number of goals scored in the total number of games played (G) subtracted by the total number of goals scored against in a total number of games played (GA)
CQF GP	Total Games played in the Conference Quarterfinals (round 1)
CQF OT GP	Total Games that went to OT in the Conference Quarterfinals
CQF W-L%	Conference Quarterfinals Total Wins divided by Total Games played in the Conference Quarterfinals (CQF GP)
CQF RS W-L%	Regular Season record between the two teams playing each other in the Conference Quarterfinals
CSF GP	Total Games played in the Conference Semifinals (round 2)
CSF OT GP	Total Games that went to OT in the Conference Semifinals
CSF W-L%	Conference Semi-Finals Total Wins divided by Total Games played in the Conference Semifinals (CSF GP)
CSF RS W-L%	Regular Season record between the two teams playing each other in the Conference Semifinals
CF GP	Total Games played in the Conference Finals (round 3)
CF OT GP	Total Games that went to OT in the Conference Finals

CF W-L%	Conference Finals Total Wins divided by Total Games played in the Conference Finals (CF GP)
CF RS W-L%	Regular Season record between the two teams playing each other in the Conference Finals
F GP	Total Games played in Finals (round 4)
F OT GP	Total Games that went to OT in the Finals
F W-L%	Finals Total Wins divided by Total Games played in the Finals (F GP)
F RS W-L%	Regular Season record between the two teams playing each other in the Conference Semifinals
ELITE F	Point system for the total number of Elite Forwards a team has. In order to be an Elite Forward: the player must be within the top 50 forwards in all three categories: Points, +/-, GP + Shooting percentage in the Total games played (GP). If a team has a player in all 3 categories team gets 1 point
ELITE D	Point system for the total number of Elite Defensemen a team has. In order to be an Elite Defensemen: the player must be within the top 50 Defensemen in four of the five categories: Points, +/-, Most Block Shots per Game, Most Takeaways, Least Giveaways in the Total games played (GP). If the team has a player in 4 of the 5 categories team gets 1 point
ELITE GK	Point system for the total number of Elite Goalies a team has. In order to be an Elite Goalie: the player must be within the top 8 Goalie in 2 categories: Games Played and Save percentage in the Total games played (GP). If a team has a player in both categories team gets 1 point
Exp	Point system to determine the playoff experience of a team. For every previous year, the team made the playoffs the team would receive one point. The parameter takes into account the previous two seasons making the maximum amount of points a team can receive 2.
AvAge	The average age of players on a team.

Unique Variables of the NBA model

ELITE.DEF	Point system for the total number of Elite Defensemen a team has. In order to be an Elite Defensemen: the player must be within the top 50 Defensemen in three of four categories: Blocks, Steals, Defensive Rebounds and Offensive Rebounds
ELITE.OFF	Point system for the total number of Elite Forwards a team has. In order to be an Elite Forward: the player must be within the top 50 forwards in all three categories: Points, Assists, Rebounds, and Field Goal percentage
X3P	Total percentage of points obtained by 3 point shots
X2P	Total percentage of points obtained by 2 point shots
FT	Free Throw Percentage

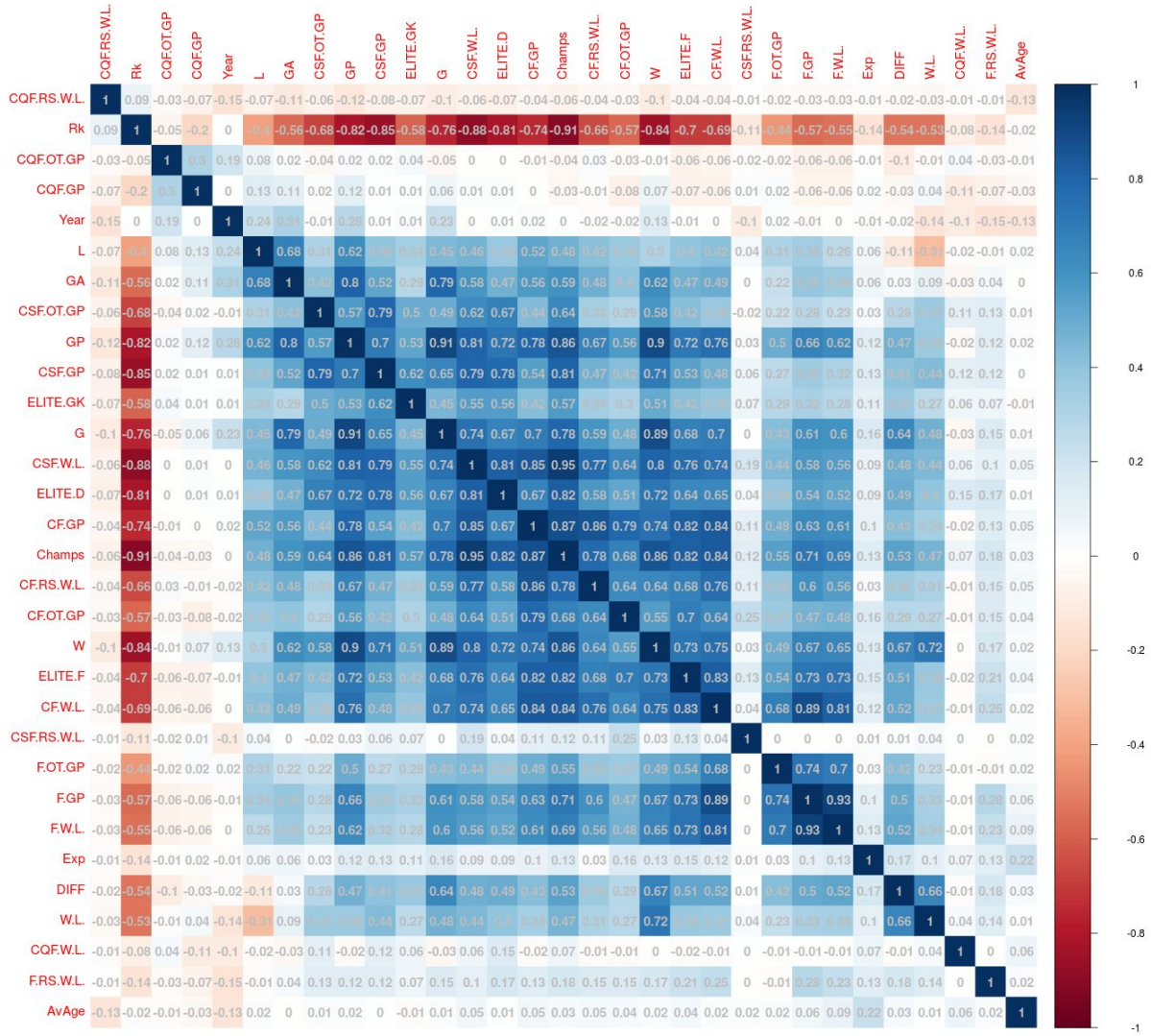
NHL

NHL Model Analysis

```
## 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 would randomly choose 50% of the data from the original data set. We used logistic regression to get a better understanding of the model to determine which variables have an influence on the model. After performing backward elimination and using our intuition, several variables were considered insignificant and were dropped. Variables such as F.W.L(Finals Win Percentage) were dropped based on our intuition 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 Difference) was dropped because it was insignificant to the model and this can be further proven from the scatter plots below. When we observe both scatter plots we can see the better the players a team has, the higher the Goal Difference. If we were to look at the scatter plot observing the number of ELITE.F (Elite Forwards) a team 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 increases, 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 an 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 proven from observing the box-plots below. As we can see the best team Rk-1 has the most number of Elite Forwards. Whereas for Elite Defenders we can see the best team Rk-1 has a lot of Elite Defenders 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.

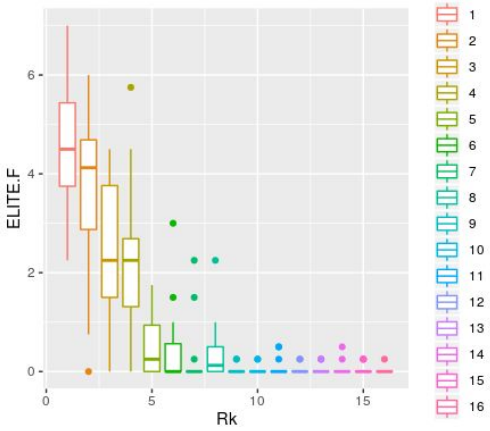
Correlation Plot (NHL)



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(Finals Win Percentage)-0.93, GP(Games Played)-0.91, F.GP(Number of Final Games Played) and G-0.89(Goals Scored), ELITE.F(Elite Forwards)-0.83, ELITE.D(Elite Defenders)-0.81, and ELITE.GK(Elite Goalies)-0.49. We may need to drop some of these variables when creating the model.

Box-Plots (NHL)

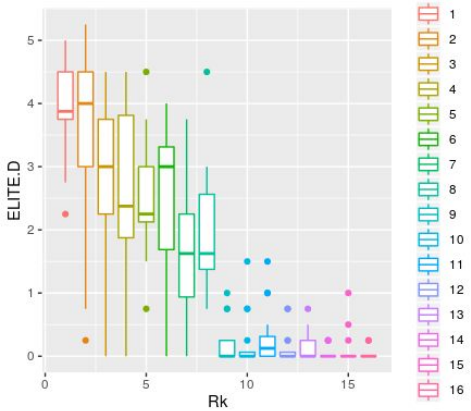
```
ggplot(NHLALL,aes(x=Rk,y=ELITE.F,colour=as.factor(Rk)))+geom_boxplot()
```



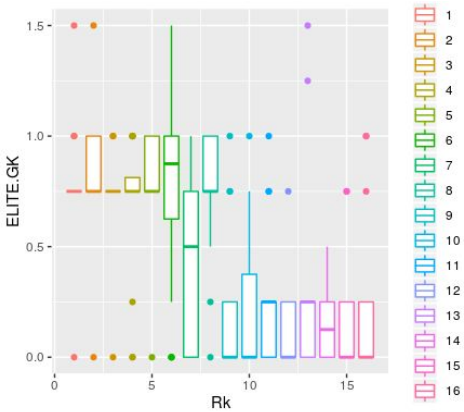
As we see in the box-plot, the x-axis is the Variable Rk where it ranges from Rk-1 which is the team that won the Championship till Rk-16, the team that did the worst. The y-axis indicates the number of Elite forwards each team has and it is evident that there are outliers in teams ranked 2,4,6,7,8,9,10,11,12,13,14,15, and 16. It is also important to note that the teams with the most number of Elite Forwards tend to be closer to winning the championship. In this box-plot, it is clearly 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 finished one spot above Rk-8.

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

As we see in the box-plot, 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 to winning the championship. However, in this box-plot, it is clearly 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 the 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()
```

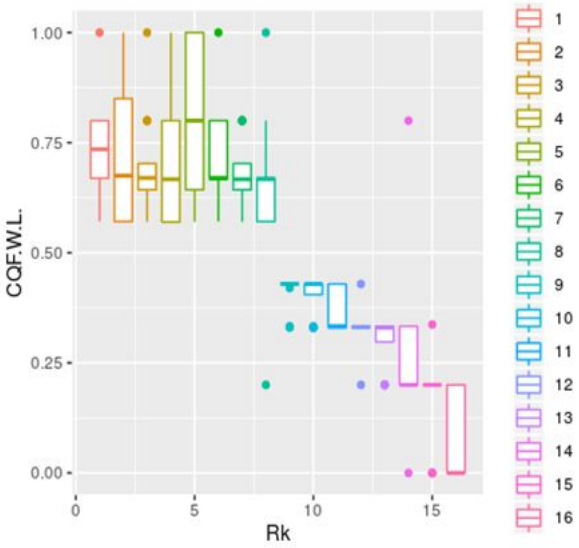


As we see in the box-plot, the x-axis is the Variable Rk where it ranges from Rk-1 which is the team that won the Championship till Rk-16, the team that did the worst. The y-axis indicates the number of Elite Goalies each team has and it is evident that there are outliers in all the teams It is also important to note that the teams with the most number of Elite Goalies tend to be closer in winning the championship. However, in this box-plot, it is clearly 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 three box-plots above, we can see that all three positions are important. However, ELITE.F (Elite Forwards) is the reason why teams win 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 than 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. Although, you do not need to have the most. Having Elite Defenders is crucial 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 championship has the most number of Elite Forwards of that season compared to the other 15 teams in the playoffs. Our model has also indicated this as well.

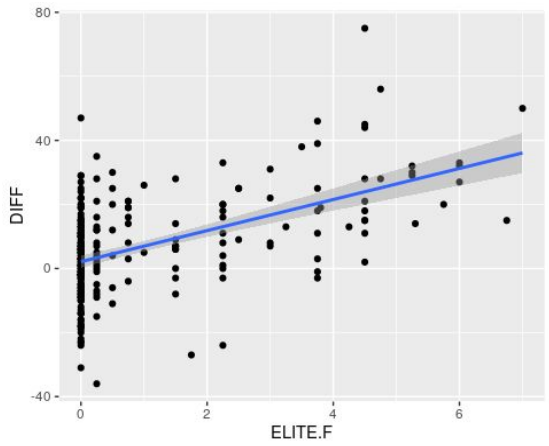
```
ggplot(NHLALL,aes(x=Rk,y=CQF.W.L,colour=as.factor(Rk)))+geom_boxplot()
```

As we see in the box-plot, 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 Quarterfinals. It is evident that there are outliers in all the teams rankings except teams ranked, Rk-4,Rk-5,Rk-11 and Rk-16. Based on intuition 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 focusing on the top eight 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.



Scatter-Plots (NHL)

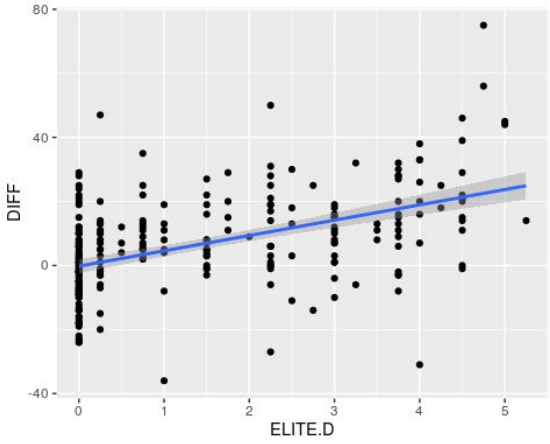
```
ggplot(NHLALL,aes(x=ELITE.F,y=DIFF))+geom_point()+geom_smooth(method="lm")
```



From observing the scatter plot which is showing the effect that Elite Forwards have on the Goal Difference, it shows the obvious. The more Elite Forwards a team has, the more goals will be scored which will result in teams having a higher Goal Difference. If we were to also observe the regression line,it clearly indicates, as the number of Elite Forward increases, 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 the Goal Difference, it shows the obvious. The more Elite Defenders a team has, fewer goals will be scored against a team, which will then 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.F(Elite Forwards) a team 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 increases, 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 higher slope, 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 than Elite Defenders.

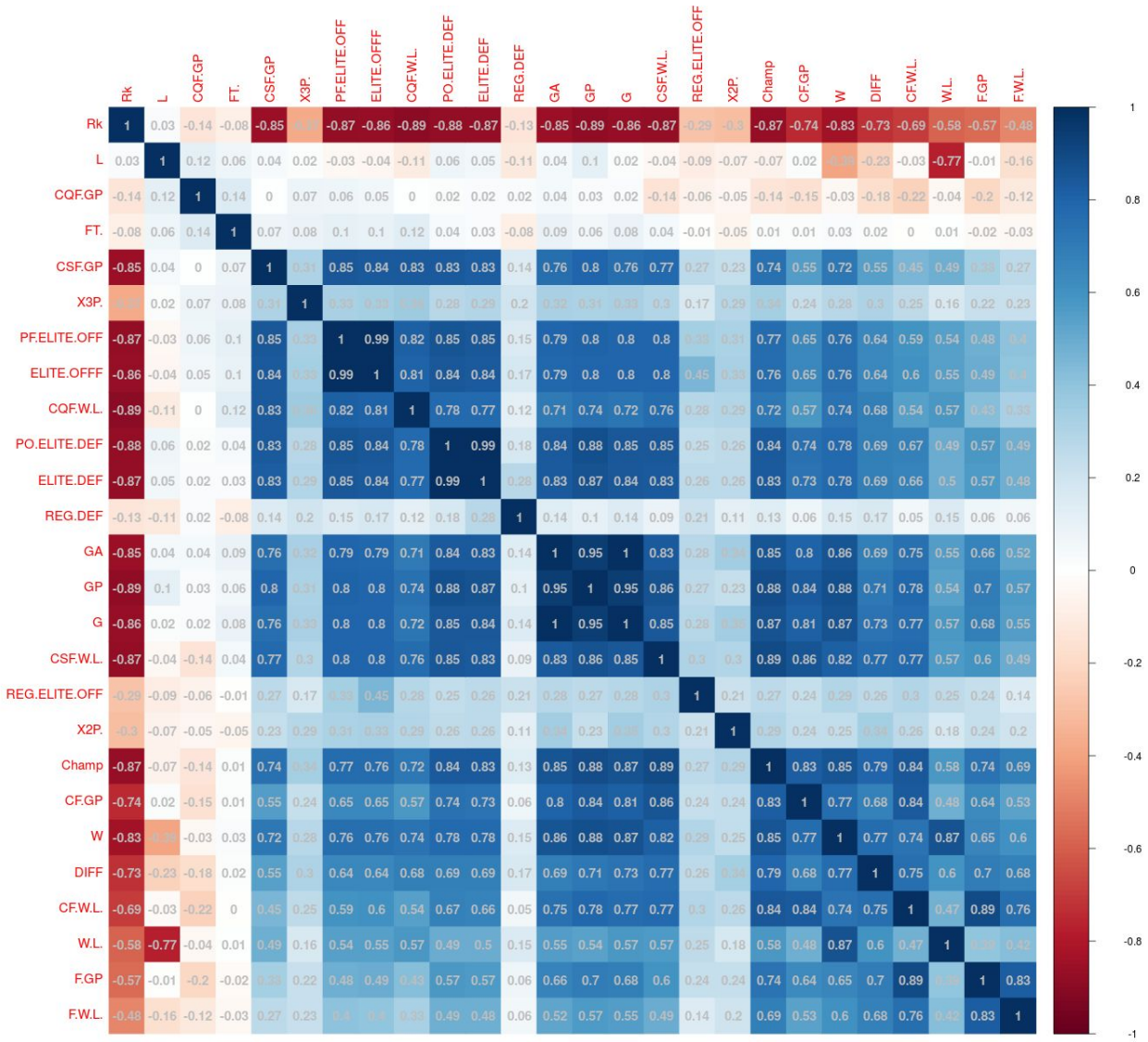
NBA

NBA Model Analysis

```
 ## Model:
## ordered(Champ) ~ ELITE.DEF + ELITE.OFF + X3P.
##                   Df    AIC    LRT  Pr(>Chi)
## <none>               117.72
## ELITE.DEF.SCORE  1 151.70 35.984 1.990e-09 ***
## ELITE.OFF.SCORE  1 123.25  7.539  0.006038 **
## X3P.              1 133.40 17.686 2.605e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The dataset used to create this model contains a total of 15 NBA seasons from year 2003 to 2018. This excludes 2012 as the playoffs were cancelled that year. Each season, 16 teams are selected to the playoffs based on performance, thus we have a dataset of size 240. Logistic regression was used to model the dataset to understand which variables had the most impact on the playoff results. We gave each team a variable between three and zero based on their performance in the playoffs; a variable defined as "Champ". The teams that won the playoffs received three, and a value of two was given to the teams that won the conference finals. A value of one was assigned to the teams that won that conference semifinals. All the other teams received a value of zero. The training dataset was created using half of the original dataset which had been randomly selected. After performing backwards elimination and using our intuition several variables were considered insignificant and were dropped. We dropped variables such as final win-loss percentage (F.W.L.) and conference semi-final win-loss percentage (CSF.W.L.), as they offered no significant understanding as to what factors led the teams to win. Hence we chose the model: polr(formula = ordered(Champ) ~ ELITE.DEF + ELITE.OFF + X3P. From this model we can see that both elite defense and offensive players are very influential to a teams chances of winning the NBA playoffs. Three point shot percentage makes the difference when it comes to predicting the champion while two point shot percentage and free throws percentage does not provide any significance when predicting whom would be the winner between the playoff teams.

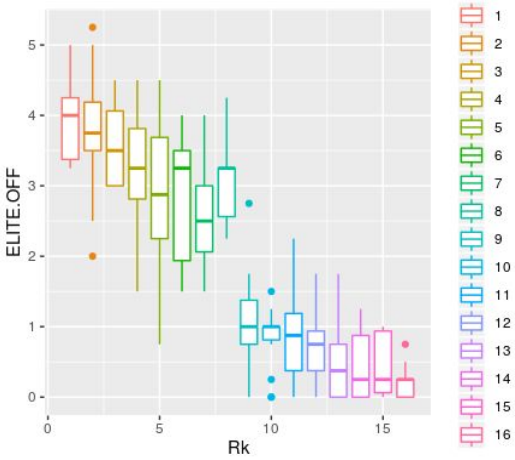
Correlation Plot (NBA)



The correlation plot indicates high correlation between elite player variables (ELITE.DEF, ELITE.OFF) and the team's goal difference, along with how well they do in all the Conference Finals and Finals. Some of these variables must be dropped to avoid multicollinearity in the final model. three-point shot percentage, two-point shot percentage, and free throw percentage do not correlate much with any other variables in the dataset.

Box Plots (NBA)

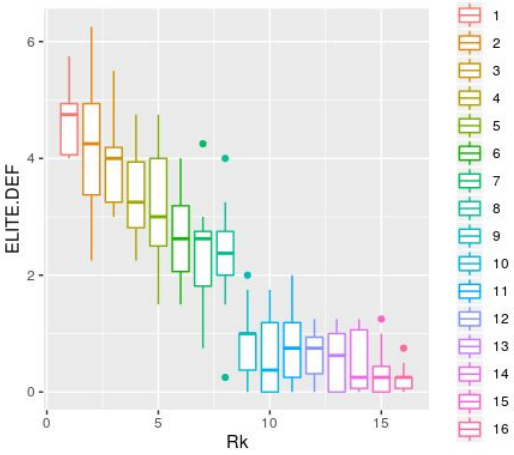
```
ggplot(nbaAllDf,aes(x=Rk,y=ELITE.OFF,colour=as.factor(Rk)))+geom_boxplot()
```



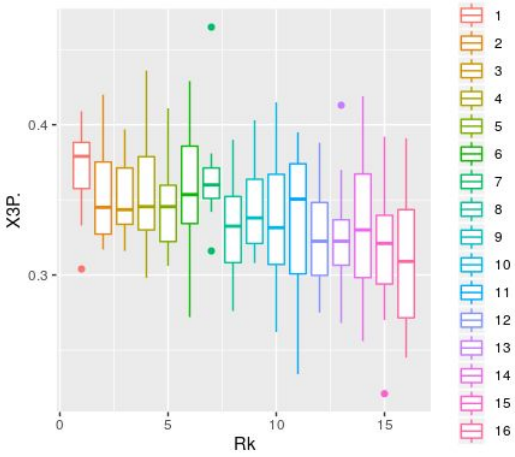
In this plot, the x-axis represents the rank of the team. Rank 1 represents the team that did the best in the playoffs, and Rank 16 represents the team that did the worst. The y-axis elite offensive parameter is calculated using the total number of elite forward players for the regular season after the trade deadline and total elite offensive players during the playoffs. The teams that made it past the conference semifinals have significantly more elite players then the teams that only made it till the conference first round. There are outliers presents in Ranks 2,8,9, 15 and 16.

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

In this plot, the x-axis represents the rank of the team as explained previously. The y-axis represents elite defensive score which is calculated using the total number of elite defensive players for the regular season after the trade deadline and the total elite defensive players during the playoffs. There are outliers present for ranks 6,7,8,15 and 16. Overall, we see a trend where the further a team makes it in the playoffs, the more number of elite players they have.




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

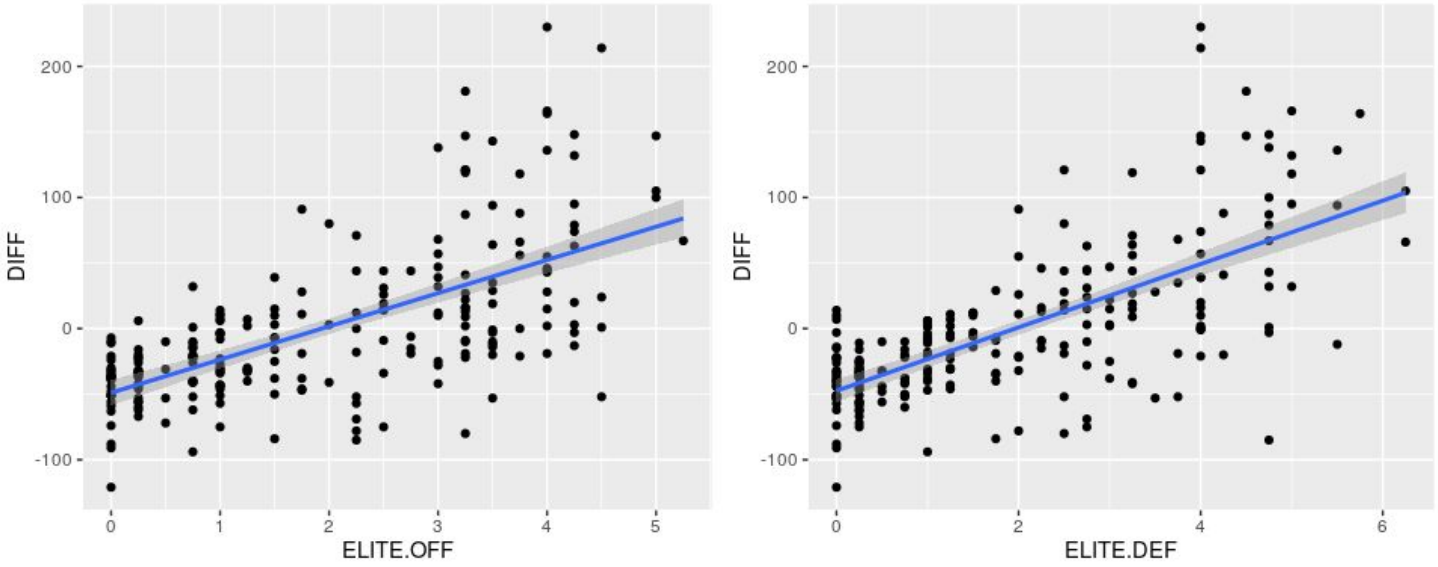


In this plot, the x-axis represents the rank of the team as explained previously. Y-axis represent the percentage of three-point shots out of all the points scored. There are a few outliers present for the 1st, 7th, 13th, and 14th ranks, however the overall trend shows that the teams that rank higher score more three-point shots.

Scatter Plot (NBA)

```
ggplot(nbaAllDf,aes(x=ELITE.DEF,y=DIFF))+geom_point()+geom_smooth(method="lm”
)

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



These two plots indicate that the major point difference between teams positively correlates with the elite defensive and offensive players that are present on the team. However, we can see from these graphs that increase of elite defensive players tend to create more of a goal difference then increase in elite offensive players.

Conclusion

Using the 30 explanatory variables, we used statistical analysis tools such as backwards and forwards elimination along with significant codes provided by R to determined the most influential variables. R provided significant codes by using p values to determine if a variable was playing an important role in the model or not. For example a variable would be very significant is it's $p < 0.001$ and it would be given the code " *** ", or is the $p < 0.01$ it would be assigned code "**" showing it is less significant than the previous variable but still important to the model. In the end only three of the 30 variables showed importance to both sports.

In the NHL, a team must have Elite Defense, Elite Forward, and have a high Conference Quarter Final Win Percentage in order to give themselves the best opportunity to win the Stanley Cup. Both explanatory variables, Elite Defense and Elite Forwards, are comprised of a multitude of general statistics collected throughout the game. For this paper, a defenseman or forward were deemed elite if they were highly skilled in numerous different areas in the game. Nevertheless, the model built showed in the game of hockey, Elite Forwards were more important for a team to have than Elite Defence. Although Elite Forwards are crucial for an NHL team to advance to the next round, Elite Defensemen help a team get to the playoffs and enhance the importance of goal-scores such as Elite Forwards.

On the other hand, the three most significant parameters in the NBA were Elite Defense, Elite Offense, and 3-point percentage. Similar to the NHL, the explanatory variables Elite Defense and Elite Offense are both composed of numerous different game-changing statistics collected by the league. Both Elite Defense and Elite Offense positively correlates with Goal Differential, indicating that elite players have a significant impact on a teams ability to win.

Both models had used very similar parameters to analyze and eliminate. The majority of parameters had been the same amongst the two models, however for the elite parameters, the NHL and NBA models had been comprised of different statistics which generally lead to winning a game. It can be seen via the models that though both the NBA and NHL require good offense and good defense, the NHL model suggests that a stronger offense is needed to have a higher chance in winning the championship, while in the NBA, the model suggests that a stronger defensive play is more important and critical in having a higher chance to win a championship.

There are limitations to further compare both sports as they both collect data differently and also analysis very different data variables. For instance, X3P, total percentage of points obtained by 3 point shots, is unique to the NBA and can not be used in the NHL. Moreover, in the NHL, a parameter called Elite Goalies however, there are no goalies in the NBA. The R-code along with the models constructed are designed to build a model for both sports because they are similar in regard to the statistics collected. As a result, the models created can not be used for other sports such as NFL or MLS, but the models for other sports would be similar as ordered logistic regression will provide the best result.

Works Cited

Analysis. (n.d.). Retrieved September 15, 2019, from http://www.nhl.com/stats/.

Basketball Statistics and History. (n.d.). Retrieved September 15, 2019, from https://www.basketball-reference.com/.

NBA Stats. (n.d.). Retrieved October 18, 2019, from https://stats.nba.com/.

NHL Stats, History, Scores, & Records. (n.d.). Retrieved September 23, 2019, from https://www.hockey-reference.com/