Goal Creation Analysis in NHL

Takehiro Matsuzawa

Stat 91r

Abstract

Goal creation in ice hockey is complicated and difficult to understand. Unlike baseball, players and puck are always moving and a combination of 5 players produces a goal. A team is always looking for a player who can contribute to team goals. National Hockey League has about 7000 goals per season. This study analyzes goals and assists in NHL to predict the contribution of each player to team goals in the following season with weighted multiple linear regression. I scraped NHL game documents from NHL official website to explore which factors are most highly associated with a team's goals next season. I divide the data into different manpower situations and different positions. Usually, the most important variables to predict the number of goals in the following seasons are the team goals (goals with assists).

I. Introduction

Hockey has different types of goals and assists. In order to investigate goals and assists thoroughly, we need to understand the types of goals and assists because some goals are produced by an effort of a team and some goals are produced solely by an effort of a single player.
Unassisted goals can be considered largely as a result of one player and assisted goals can be considered largely as a result of players with a goal or assists.

I define the performance of player as the number of goals a team can score while the player is on the ice. In this paper we predict the number of goals a team can score while a player is on the ice in the following season, not the number of assists or goals he can directly produce in the following season because I examine the contribution of a player to the team.

Today sports analysts are not sure about explanatory variables for predicting goals of a team next year. For instance, some people use only goals and 1st assists to evaluate player performance, but other people use points (goals + 1st assists +2nd assists) to evaluate players.

In this study, we examine all kinds of goals and assists to predict the number of goals a team can score while he is on the ice.

For example, this study can answer:

1) What are the effects of 1st-Assisted Goals for team goal prediction?
2) Should you acquire a player who has more Non-Assisted Goals or 1st-Assisted Goals?
3) Which variable is the most important explanatory variable for evaluating the contribution of a player to the team offense?

To examine these questions, I scraped the data of four seasons on the NHL website. I used a data set which consists of goals and assists for the past four seasons. The NHL has all the goals, assists and players on the ice when a goal is scored. I used this large dataset to find out the variables that might help teams in NHL to pinpoint what kind of players they should acquire in order to enhance the team's offensive performance.

II. Data Scraping

1. Goal Information

I downloaded the goal information data from the NHL official website. The following link is an example for the data source. The last four digits are the id for a game. It ranges from 0001 to 1230 because each season has 1230 games except 2012-2013 season. It only ranges from 0001 to 0720 in 2012-2013 season because it only has 720 games.
(*http://www.nhl.com/scores/htmlreports/20142015/GS020036.HTM*)
20142015 represents the season. If you change this number, you can get the information of another season.

I categorized goals into 1st Assisted Goals, 2nd Assisted Goals and Non-assisted Goals. Each goal has different variables. "Players" mean players on the ice who did not score nor assist. For example, 1stAssistPlayers are players on the ice when a 1st-Assisted Goal is scored, but they did not score nor assist.

1. Divide the situations into equal handed situations and power play situations. Short-handed goals are too few to think. Thus I decided to remove them.
2. 1st Assisted Goals have three variables, 1stAssistGoal, 1stAssist and 1stAssistPlayers.
3. 2nd Assisted Goals have three variables, 2ndAssistGoal, 2ndAssist,1stAssist and 2ndAssistPlayers
4. Non-assisted Goals have two variables, NonAssistGoal and NonAssistPlayers.
5. **I removed empty net goals and shootout goals** because they do not show a player's performance to a team at games.

The following is the summary table.

Variable Type	1st Assisted Goals	2nd Assisted Goals	Non-assisted Goals
Goal	1stAssistGoal	2ndAssistGoal	NonAssistGoal
Assist	1stAssist	1stAssist	
		2ndAssist	
Players on The Ice	1stAssistPlayers	2ndAssistPlayers	NonAssistPlayers

2. Player Information

I also scraped each club's home page to get the data of uniform numbers which are necessary to specify players based on numbers and teams.
Example: Boston Bruins page looks like (http://bruins.nhl.com/club/stats.htm).

The following is an example table of information I made by scraping the NHL website and NHL team websites.

Name	Year	Team	Number	NonAssistGoal	1stAssistGoal	2ndAssistGoal	1stAssist	2ndAssisted1st	2ndAssist	NonAssistPlay	1stAssistPlaye	2ndAssistPlaye
James Reimer	20132014	TOR	34	0	0	0	0	0	0	1	11	23
Tyler Bozak	20132014	TOR	42	0	4	7	6	12	9	2	7	8
Jake Gardiner	20132014	TOR	51	0	2	5	1	4	7	1	12	19
Phil Kessel	20132014	TOR	81	0	9	16	5	14	9	2	5	9
Carey Price	20132014	MTL	31	0	0	0	0	0	1	6	27	66
Francis Bouillo	20132014	MTL	55	0	0	2	1	0	2	2	11	10
Raphael Diaz	20132014	MTL	61	0	0	0	2	5	1	0	3	3
Lars Eller	20132014	MTL	81	1	2	5	1	6	5	1	1	5
Carl Gunnarsso	20132014	TOR	36	1	0	2	0	5	7	3	17	18

I stored the data as csv format in order to use it in RStudio later.

3. Time Information

I downloaded the time information data for the past four seasons from the war-on-the-ice website (http://war-on-ice.com/).
The data on war-on-the-ice website gives us each player's information such as time on the ice, the number of assists and salary. I only used the information about the total time on the ice because assists are not divided into 1st assists or 2nd assists.

The following is an example of information on war-on-the-ice website.

Name	pos	Team	Gm	season	Age	Salary	AAV	HIT	HIT-	A1	A2	SH
Aaron.Ekblad	D	FLA	81	20142015	18	0.925	0.925	98	123	8	8	109
Aaron.Volpatti	L	WSH	2	20142015	29	0.6	0.575	3	2	0	0	1
Adam.Burish	C	S.J	20	20142015	31	1.85	1.85	11	25	2	0	19
Adam.Clendeni	D	CHI/VAN	21	20142015	21	0.818	0.743	9	35	0	1	14
Adam.Cracknel	C	CBJ	17	20142015	29	0.6	0.6	42	21	1	0	15
Adam.Henrique	CL	N.J	75	20142015	24	3	4	54	70	7	6	93
Adam.Larsson	D	N.J	64	20142015	21	0.9	0.9	78	92	8	9	79

I stored the data as csv format for data analysis in RStudio.

III. Data Cleaning

Data Combine

I made a huge data set by combining time information and goal information based on players' names. If a player transfers to another team during the regular season, I combine the player's data during one season.

For example, Ryan Callahan played for New York Rangers and Tampa Bay Lightnings in 2013-2014 season. I merged two rows of Ryan Callahan into one row.

Previous Data

Ryan Carter	20132014	N.J	20	0	1	6	0	1	1	0	2	1
Ryan Callahan	20132014	NYR	24	0	2	4	0	8	2	2	2	3
Ryan Callahan	20132014	T.B	24	0	1	2	0	1	3	2	3	3
Ron Hainsey	20132014	CAR	65	1	1	2	2	3	5	4	9	17
Roman Polak	20132014	STL	46	0	0	4	0	2	6	2	9	10
Roman Josi	20132014	NSH	59	0	2	8	3	7	5	2	9	25
Roman Horak	20132014	EDM	51	0	0	0	0	1	0	0	2	2
Robyn Regehr	20132014	L.A	44	0	0	3	1	4	6	3	4	16
Robin Lehner	20132014	OTT	40	0	0	0	0	0	0	5	17	34

New Data

Ryan Carter	20132014	N.J	20	0	1	6	0	1	1	0	2	1
Ryan Callahan	20132014	NYR	48	0	3	6	0	9	5	4	5	6
Ron Hainsey	20132014	CAR	65	1	1	2	2	3	5	4	9	17
Roman Polak	20132014	STL	46	0	0	4	0	2	6	2	9	10
Roman Josi	20132014	NSH	59	0	2	8	3	7	5	2	9	25
Roman Horak	20132014	EDM	51	0	0	0	0	1	0	0	2	2
Robyn Regehr	20132014	L.A	44	0	0	3	1	4	6	3	4	16
Robin Lehner	20132014	OTT	40	0	0	0	0	0	0	5	17	34

Player Selection

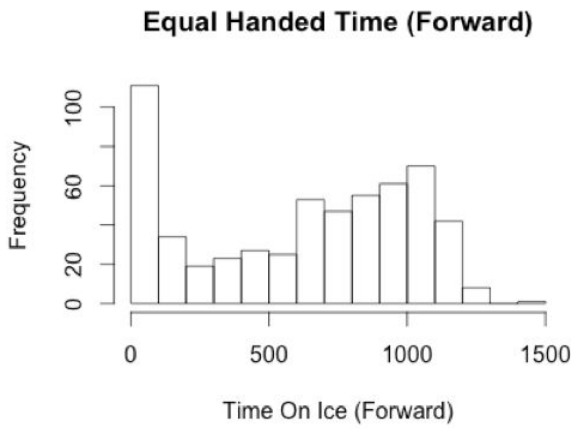
For each different manpower situation and position, I've decided to remove players who play fewer than certain minutes because the information of players who played more minutes is more reliable.

Player A scores one goal and he only plays for 40 minutes in the entire season.
Player B scores 40 goal and plays 1600 minutes in the entire season.
Player A and Player B score the same number of goals per minute.
However, player B is much more reliable to predict the performance in the following reason because he plays for more minutes.

I decided to plot the time on the ice of the entire 2014-2015 season to see distribution of players in terms of time on the ice.

Equal Handed Situations (2014-2015 season)

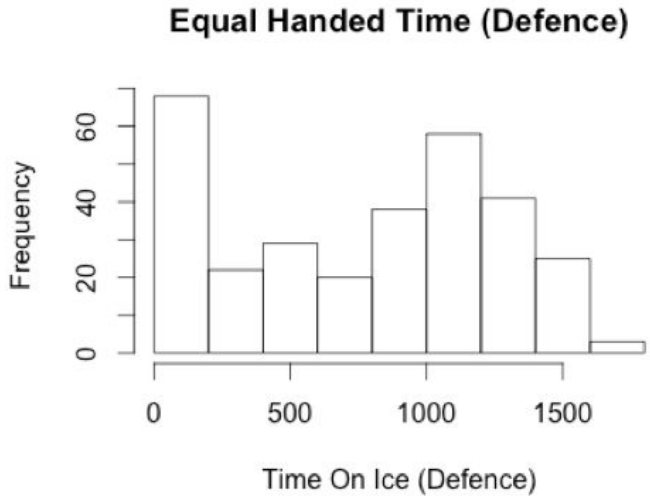
Forwards



For forwards in equal-handed situations, I have decided to remove players who play fewer than 500 minutes because players who play more than 500 minutes look close to a normal distribution. The other years are in the Appendix A. Also 500 minutes is about 25 games. I think it makes sense to remove players who play fewer than 30 percent of games in the entire season. The data of other seasons have similar distribution.

In the 2012-2013 season, there are only 48 games due to a strike and I have decided to only consider players who play more than 292 (500*48/82) minutes.

Defensemen

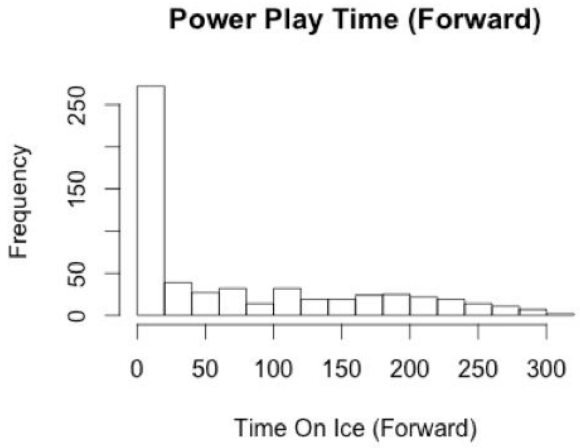


For defensemen in equal-handed situations, I have decided to remove players who play less than 500 minutes because players who play more than 500 minutes look like a normal distribution. The other years are in the Appendix A. The data of other seasons have similar distribution.

In the 2012-2013 season, there are only 48 games due to a strike and I have decided to consider players who play more than 292 minutes.

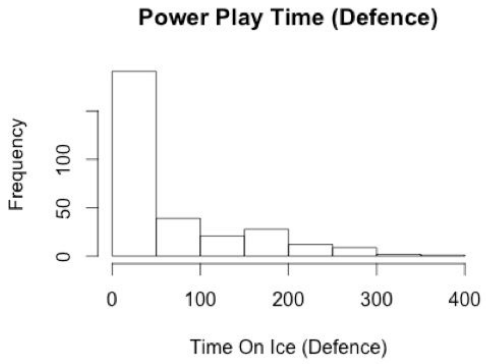
Power Play Situations (2014-2015 season)

Forwards



In power play situations, I have decided to remove players who play less than 50 minutes because the graph is extremely left-skewed and the distribution of players who play longer than 50 minutes looks close to a normal distribution. The other years' are in the Appendix A. The data of other seasons have similar distributions.
In the 2012-2013 season, there are only 48 games due to a strike and I have decided to consider players who play more than 58 (100*48/82) minutes.

Defensemen



In power play situations, I have decided to remove players who play fewer than 100 minutes because players who play longer than 100 minutes look close to a normal distribution. The other years' are in the Appendix A. The data of other seasons have similar distributions.

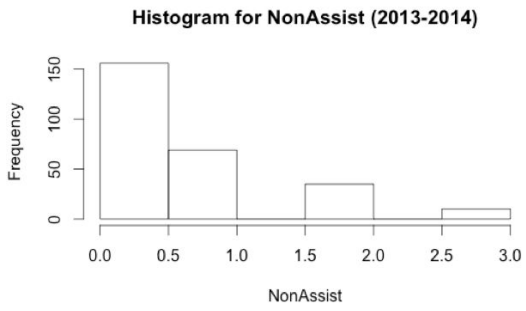
In the 2012-2013 season, there are only 48 games due to a strike and I have decided to consider players who play more than 58 minutes.

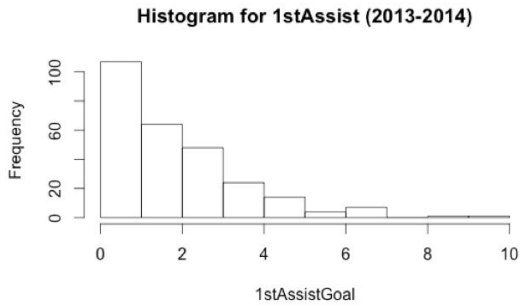
IV. Analysis of Variables

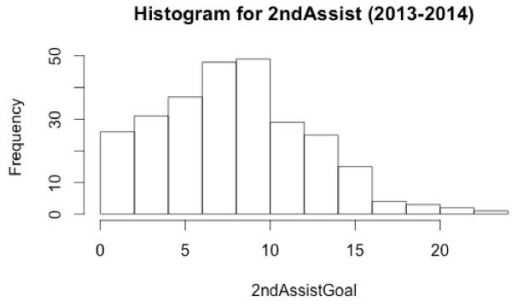
Variable Transformation

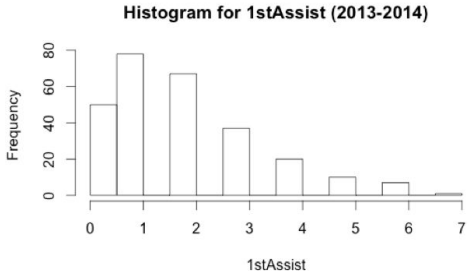
A multiple linear regression assumes the normal distribution of each variable. I need to look at the distribution of each variable.

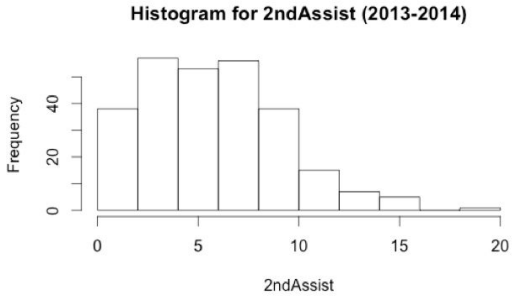
Explanatory Variables

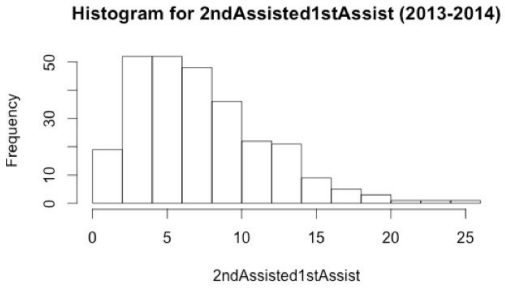


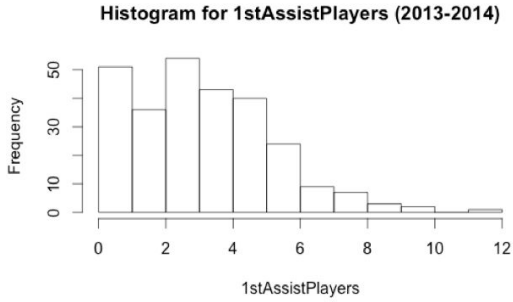


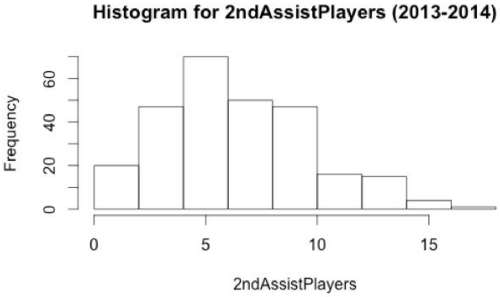


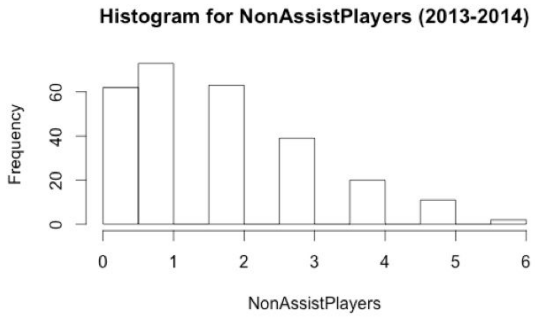




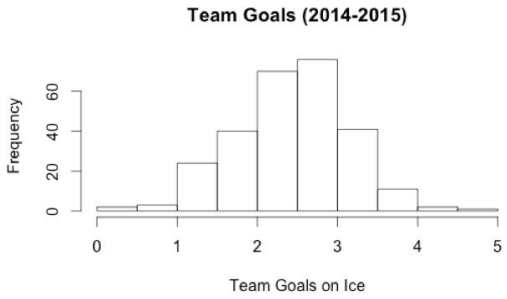








Dependent Variable



All the independent variables are right skewed. We need to transform variables with log-transformation or square root transformation. I have decided to choose square root transformation because there are many players whose values are close to 0 and log 0 is -infinity.

The dependent variable (team goals) looks normally distributed. Thus we do not need to transform the team goals.

***All the variables below are square rooted.**

V. Analysis of Linear Models

I assumed there is a linear relationship between the number of goals a team would score in the following year and explanatory variables this year. I used the weighted linear regression instead of just using the linear regression. I weighted the regression on time because players who played longer are more reliable.
I decided to divide players into defensemen and forwards and divide situations into equal-handed situations and power play situations.

I think about the following four cases. The red letters describe the four situations.

	Forward	Defensemen
Equal-Handed	Equal-Handed (Forwards)	Equal-Handed (Defensemen)
Power Play	Power Play (Forwards)	Power Play (Forwards)

In each situation I regress the number of team goals in the following season on explanatory variables this year so that I tried to determine which variable is the most significant to predict team goals in the following season.

The explanatory variables are the following.

Variable Type	1st Assisted Goals	2nd Assisted Goals	Non-assisted Goals
Goal	1stAssistGoal	2ndAssistGoal	NonAssistGoal
Assist	1stAssist	1stAssist	
		2ndAssist	
Players on The Ice	1stAssistPlayers	2ndAssistPlayers	NonAssistPlayers

First, I would like to look closely at the data of forwards in equal-handed situations.

Equal-Handed Situations (Forwards)
1. Linear Model for Each Year

The following is the result of the weighted linear regression of 2014-2015 on 2013-2014 variables. 2nd-Assisted Goals' 1st assist, 2nd-Assists and 2nd-Assisted Goals seem to be statistically significant variables to explain the number of goals a team can score when a player is on the ice.

```
Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      0.740849   0.217469   3.407 0.000762 ***
sqrt(player20132015F500$NonAssistGoal.xPer60)   -0.048088   0.251926  -0.191 0.848768
sqrt(player20132015F500$X1stAssist.xPer60)       0.228560   0.226853   1.008 0.314620
sqrt(player20132015F500$X2ndAssisted1stAssist.xPer60)  0.792474   0.222040   3.569 0.000427 ***
sqrt(player20132015F500$X2ndAssist.xPer60)       0.474268   0.229763   2.064 0.039994 *
sqrt(player20132015F500$NonAssistPlayers.xPer60) 0.060669   0.219946   0.276 0.782894
sqrt(player20132015F500$X1stAssistPlayers.xPer60)  0.254369   0.217843   1.168 0.244009
sqrt(player20132015F500$X2ndAssistPlayers.xPer60)  0.148118   0.248041   0.597 0.550927
sqrt(player20132015F500$X1stAssistGoal.xPer60)   -0.005665   0.242134  -0.023 0.981353
sqrt(player20132015F500$X2ndAssistGoal.xPer60)   0.878175   0.231061   3.801 0.000180 ***
---
```

The following is the result of the weighted linear regression of 2013-2014 on 2012-2013 variables. 2nd-Assisted Goals' 1st assist, Non-Assisted Goals and 2nd-Assisted Goals seem to be statistically significant variables to explain the number of goals a team can score when a player is on the ice.

```
Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        0.9289     0.2490   3.730 0.000238 ***
sqrt(player20122014F500$NonAssistGoal.xPer60)      0.5409     0.2524   2.143 0.033119 *
sqrt(player20122014F500$X1stAssist.xPer60)         0.2040     0.2003   1.018 0.309472
sqrt(player20122014F500$X2ndAssisted1stAssist.xPer60)  0.9386     0.2226   4.217 3.48e-05 ***
sqrt(player20122014F500$X2ndAssist.xPer60)         0.2670     0.2198   1.215 0.225552
sqrt(player20122014F500$NonAssistPlayers.xPer60)   0.2695     0.2038   1.323 0.187218
sqrt(player20122014F500$X1stAssistPlayers.xPer60)  0.2164     0.1843   1.174 0.241479
sqrt(player20122014F500$X2ndAssistPlayers.xPer60)  0.1913     0.2279   0.840 0.401930
sqrt(player20122014F500$X1stAssistGoal.xPer60)     0.1548     0.1967   0.787 0.432264
sqrt(player20122014F500$X2ndAssistGoal.xPer60)     0.4773     0.2322   2.055 0.040892 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The following is the result of the weighted regression of 2012-2013 on 2011-2012 variables. 2nd-Assisted Goals' 1st assists, 1st-Assisted Goals and 2nd-Assisted Goals seem to be statistically significant variables to explain the number of goals a team can score when a player is on the ice.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.57914	0.25037	2.313	0.02160	*
sqrt(player20112013F500$NonAssistGoal.xPer60)	0.36891	0.30401	1.213	0.22618	
sqrt(player20112013F500$X1stAssist.xPer60)	0.21448	0.25414	0.844	0.39957	
sqrt(player20112013F500$X2ndAssisted1stAssist.xPer60)	0.78097	0.26374	2.961	0.00339	**
sqrt(player20112013F500$X2ndAssist.xPer60)	0.52542	0.31402	1.673	0.09565	.
sqrt(player20112013F500$NonAssistPlayers.xPer60)	0.06404	0.26569	0.241	0.80974	
sqrt(player20112013F500$X1stAssistPlayers.xPer60)	-0.02243	0.30278	-0.074	0.94101	
sqrt(player20112013F500$X2ndAssistPlayers.xPer60)	0.39530	0.28050	1.409	0.16010	
sqrt(player20112013F500$X1stAssistGoal.xPer60)	0.64560	0.24952	2.587	0.01029	*
sqrt(player20112013F500$X2ndAssistGoal.xPer60)	0.62996	0.25601	2.461	0.01460	*

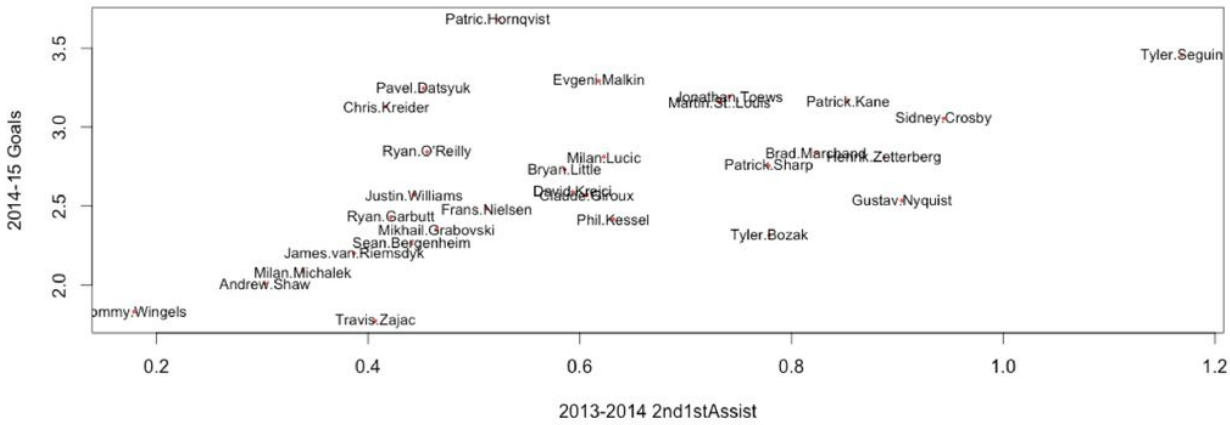
For the past three seasons, 2nd-Assisted Goals' 1st assist has been a statistically significant variable to explain the number of goals a team can score when a player is on the ice.

2.Verification of Significant Variable

In order to see whether 2nd-Assisted Goals' 1st assist is actually important, I decided to use the following method. The purpose of the following method is to hold all the variables the same except the most statistically significant variable. At equal-handed situations for forwards, the most statistically significant variable is 2nd-Assisted Goals' 1st assist.

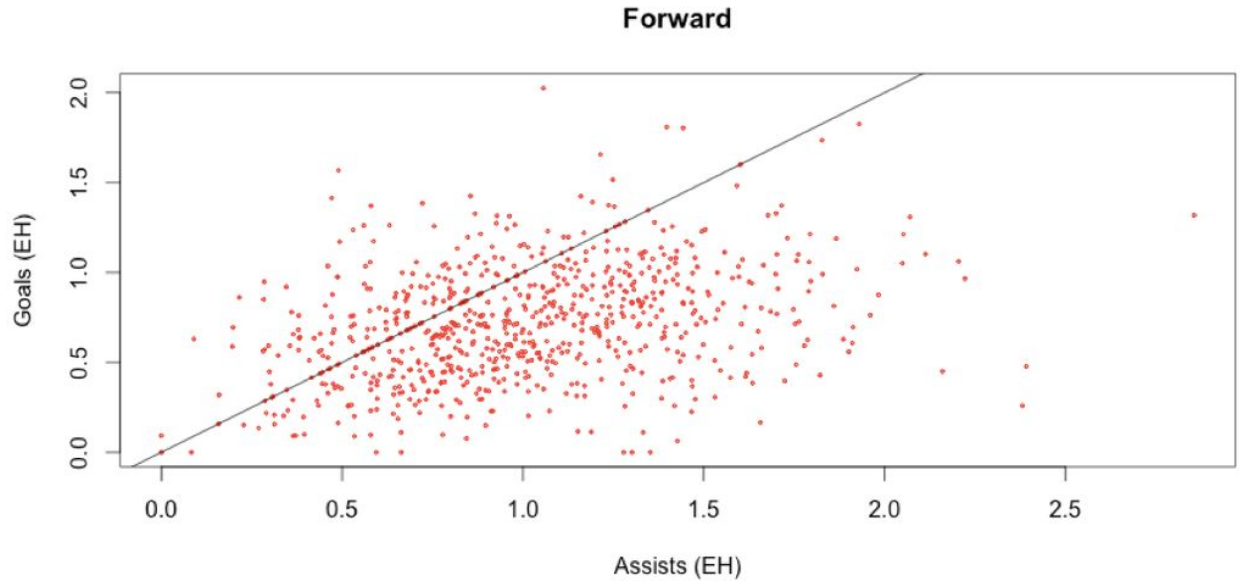
1. Predict the number of goals a team can score while he is on the ice without using the most statistically significant variable. Let this predicted number of goals as P_G.
2. Rank the players based on P_G and reorder players. By doing this, you can hold other variables constant.
3. Get the top 30 players based on P_G.
4. Predict the number of goals a team can score while he is on the ice with all variables including 2ndAssisted1stAssists for top 30 players
5. Compare the predicted goals and the actual goals. Plot them for top 30 players. Since other variables are constant, you can see the effect of 2ndAssisted1stAssists.

The following is the graph for 2013-14's 2nd-Assisted Goals' 1st assists and 2014-15 Goals for forwards. You can see a clear linear relationship between 2013-14's 2nd-Assisted Goals' 1st assists and 2014-15 Goals. There are many good players in this list. For example, Tyler Seguin, Sidney Crosby and Patrick Kane are on the list. However there are some very good players who are not on this list such as Alex Ovechkin and Steven Stamkos.



Alex Ovechkin and Steven Stamkos are some of the few players who usually score more goals than provide assists in a season. The following graph is goals vs assists per 60 minutes of forwards who play more than 500 minutes in a season between 2012 and 2015. Players above the 45-degree line indicate players who provide more goals than assists. As you can tell from the graph, you can see there are many more players who provide more assists than score goals in 60 minutes.

Goals vs Assists of Forwards in 2012-2015 season



The following is the table of goals and assists by Alex Ovechkin, Steven Stamkos and Patrick Kane.
The number left to the left slash is the number of goals. The number between the slashes is the number of assists. The number right to the right slash is the number of points. For example, in 2012-2013 season, Alex Ovechkin scored 32 goals and provided 24 assists. 2012-2013 season has only 48 games. Other seasons have 82 games.

	Alex Ovechkin	Steven Stamkos	Patrick Kane
2012-2013	32 / 24 / 56	29 / 28 / 57	23 / 32 / 55
2013-2014	51 / 28 / 79	25 / 15 / 40*	29 / 40 / 69
2014-2015	53 / 28 / 81	43 / 29 / 72	27 / 37 / 64

*Steven Stamkos played only 37 games in 2013-2014 season due to injury.

As you see on the table, they belong to different types of players from Patrick Kane. Patrick Kane belongs to a group where a player provides assists more than goals. This explains the lack of some star players like Alex Ovechkin and Steven Stamkos because they score more goals than assists.

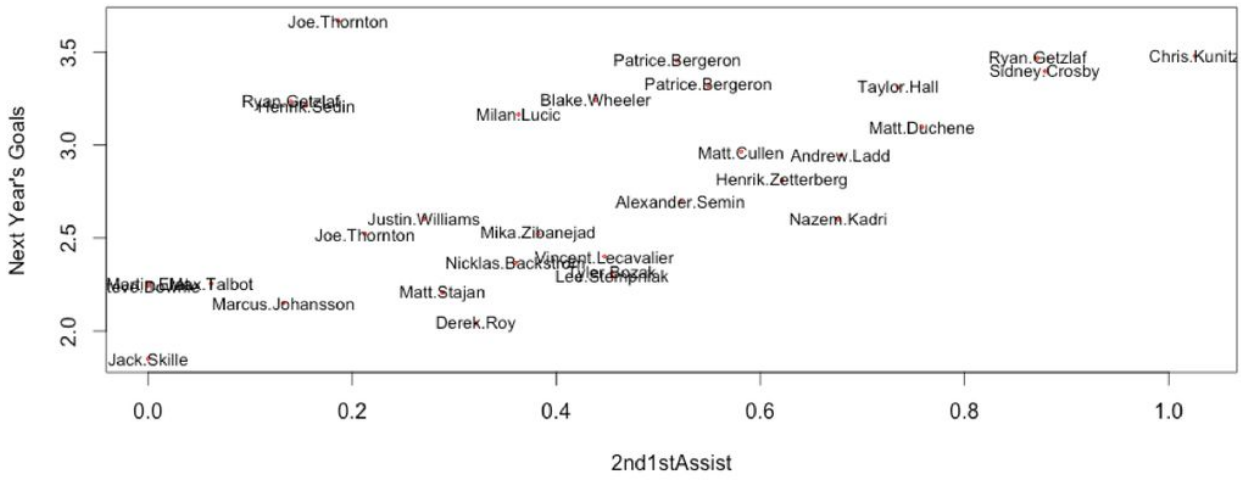
Next we look at the weighted linear model result for all the forwards who play more than 500 minutes in a season between 2011 and 2015. I combined data of each year and analyzed the whole data with RStudio.

3. Weighted Linear Model Result for 2011-2015 Season

```
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       1.42036    0.08398  16.914  < 2e-16 ***
sortPreGoalplayer20112015F500$NonAssistGoal.xPer60 0.88081    0.38149   2.309  0.02122 *
sortPreGoalplayer20112015F500$X1stAssist.xPer60    0.29929    0.22244   1.345  0.17887
sortPreGoalplayer20112015F500$X2ndAssisted1stAssist.xPer60 0.64988  0.10093  6.439 2.14e-10 ***
sortPreGoalplayer20112015F500$X2ndAssist.xPer60    0.38136    0.11567   3.297  0.00102 **
sortPreGoalplayer20112015F500$NonAssistPlayers.xPer60 0.25940  0.24564   1.056  0.29129
sortPreGoalplayer20112015F500$X1stAssistPlayers.xPer60 0.27173 0.16020   1.696  0.09026 .
sortPreGoalplayer20112015F500$X2ndAssistPlayers.xPer60 0.20262 0.11288   1.795  0.07305 .
sortPreGoalplayer20112015F500$X1stAssistGoal.xPer60 0.50224   0.20564   2.442  0.01482 *
sortPreGoalplayer20112015F500$X2ndAssistGoal.xPer60 0.49036   0.09738   5.035 5.97e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.82 on 756 degrees of freedom
Multiple R-squared:  0.216,     Adjusted R-squared:  0.2067
F-statistic: 23.14 on 9 and 756 DF,  p-value: < 2.2e-16
```

For the past three seasons, 2nd-Assisted Goals’ 1st assist is the most significant variable to predict the number of goals a team can score in a season.
Thus I decided to verify the importance of 2nd-Assisted Goals’ 1st assists as the same procedure in the above.

First, I ranked the players according to the number of goals a team can score when the player is on the ice without using the most statistically significant variable, 2ndAssisted1stAssists. The graph shows the top 30 players based on this result.
You can see a linear relationship between 2nd-Assisted Goals' 1st assists and Next Year's Goals in the graph.



Ryan Gatzlef and Sidney Crosby have many more 2nd-Assisted goals' 1st assists than Marcus Johansson even though their other variables are almost the same. Also a team scores more goals when either Ryan Gatzlef or Sidney Crosby is on the ice than when Marcus Johansson is on the ice. There is a strong relationship between 2nd-Assisted goals' 1st assists in the previous season and the number of goals a team scores in the following season.

There are some potential reasons why 2nd-Assisted Goals' 1st-Assist is the important variable.

1. 2nd-Assisted Goals' 1st Assists are result of teamwork. A player could shoot the puck when he received a pass. However he decided to pass the puck for another player for a team to score with higher probability. His sacrifice is beneficial for a team to produce more goals. This means that a player is unselfish and can play for a team.
2. 2nd-assisted Goals' 1st Assists are result of team plays. Thus this player can collaborate with other players to produce goals. Since hockey is team sports and it is difficult to produce goals by himself especially at NHL level. 2nd-assisted goals' 1st assists mean that a player can collaborate with other players to score a goal.

Next we look at defensemen at equal handed situations.

Equal Handed Situations (Defensemen)

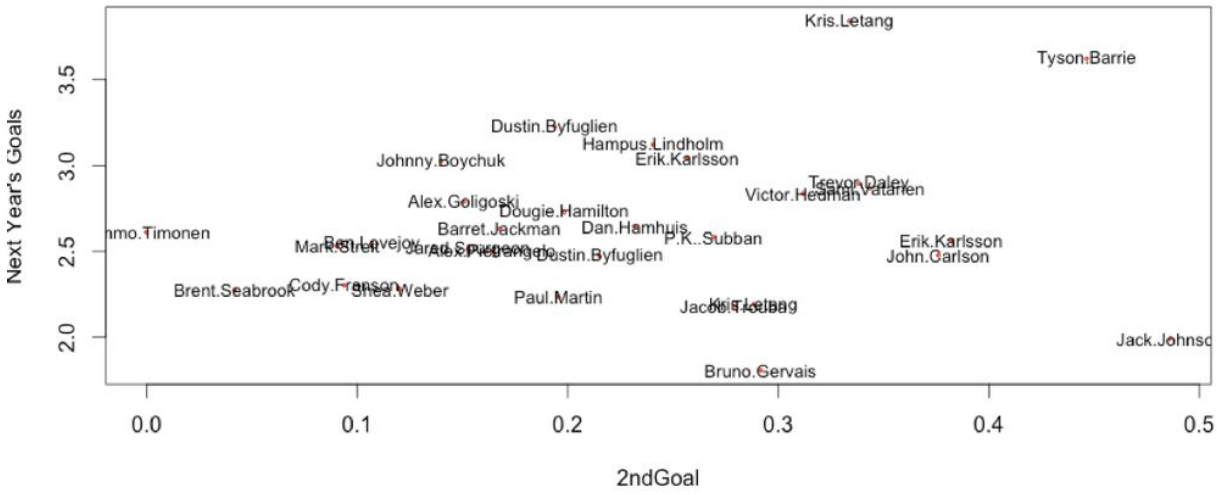
1. Weighted Linear Model Result for 2011-2015 Season

The following is based on the data between 2011 and 2015 season for defensemen.
For defensemen, 2nd-assisted goals are the most important predictor for the number of goals a team can score when he is on the ice.

```
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        1.61773    0.19537   8.280 2.55e-15 ***
sqrt(NonAssistGoal.xPer60)         0.18843    0.26772   0.704 0.481992
sqrt(X1stAssist.xPer60)            0.31915    0.16392   1.947 0.052328 .
sqrt(X2ndAssisted1stAssist.xPer60) 0.32428    0.15828   2.049 0.041224 *
sqrt(X2ndAssist.xPer60)            0.19760    0.16309   1.212 0.226480
sqrt(NonAssistPlayers.xPer60)     -0.09559    0.15395  -0.621 0.535051
sqrt(X1stAssistPlayers.xPer60)     0.11825    0.17423   0.679 0.497754
sqrt(X2ndAssistPlayers.xPer60)     0.26935    0.15956   1.688 0.092279 .
sqrt(X1stAssistGoal.xPer60)        0.24249    0.19700   1.231 0.219168
sqrt(X2ndAssistGoal.xPer60)        0.50296    0.14736   3.413 0.000717 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.7 on 354 degrees of freedom
Multiple R-squared:  0.1181,    Adjusted R-squared:  0.09567
F-statistic: 5.267 on 9 and 354 DF,  p-value: 9.541e-07
```

2.Verification of Significant Variable

First, I ranked the players according to the number of goals a team can score when the player is on the ice without using the most statistically significant variable, 2nd-assisted goals. The graph shows the top 30 players. You can see a linear relationship between 2nd-Assisted Goals' 1st assists and Next Year's Goals in the graph.



There are potentially some reasons why 2nd-Assisted goals are the most important explanatory variable. 2nd-Assisted goals' 1st-assists are important as well. Usually it takes more time for a defender to enter the offensive zone than forwards. Thus a defenseman tends to score goals when a team passes the puck in the offensive zone. As a result, defensemen usually have 2nd-Assisted goals rather than scores non-assisted goals which usually happen with a turnover. Thus, 2nd-assisted goals and 2nd-assisted goals' 1st-assists become statistically important.

Power Play Situations (Forwards)

The following analysis is based on the data from 2011 to 2015 seasons for forwards.
Power Play situations are rarer than equal-handed situations. Thus I have decided to make two sets of explanatory variables to predict the number of goals a team can score when a player is on the ice per 60 minutes.

Set 1. Only Power Play Data in the previous year
Set 2. Power Play and Equal-Handed Data in the previous year

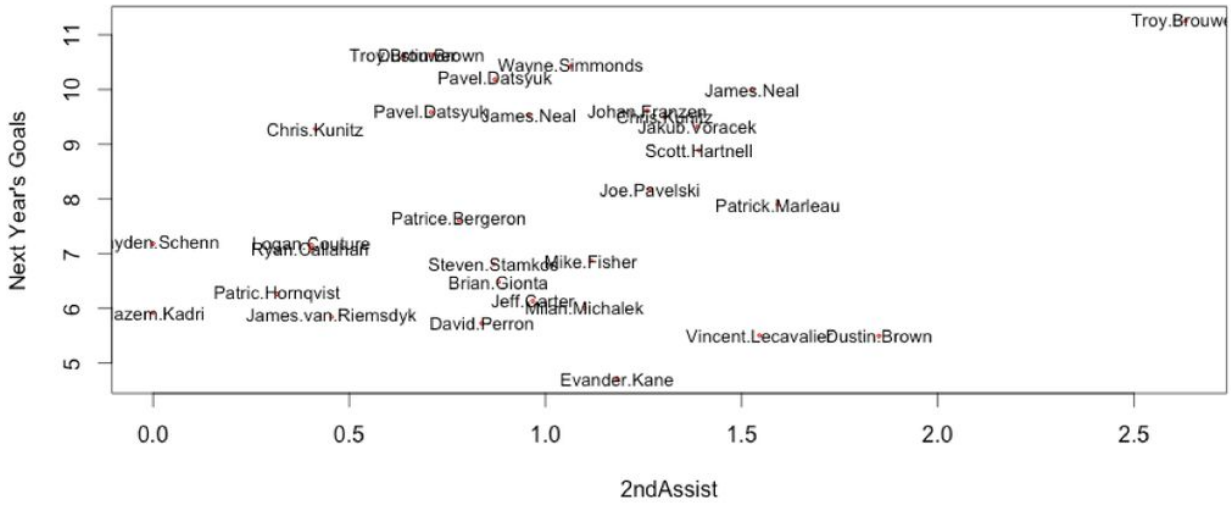
Set 1. Only Power Play Data in the previous year

Weighted Linear Model Result for 2011-2015 Season

```
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      5.04139    0.53467   9.429  < 2e-16 ***
sqrt(player20112015PPF100$NonAssistGoal.xPer60) -1.63628    0.82002  -1.995 0.046690 *
sqrt(player20112015PPF100$X1stAssist.xPer60)    -0.08298    0.36062  -0.230 0.818136
sqrt(player20112015PPF100$X2ndAssisted1stAssist.xPer60) -0.07391    0.22907  -0.323 0.747138
sqrt(player20112015PPF100$X2ndAssist.xPer60)     0.81987    0.22901   3.580 0.000387 ***
sqrt(player20112015PPF100$NonAssistPlayers.xPer60)  0.10973    0.43044   0.255 0.798909
sqrt(player20112015PPF100$X1stAssistPlayers.xPer60) -0.11891    0.26965  -0.441 0.659465
sqrt(player20112015PPF100$X2ndAssistPlayers.xPer60)  0.65871    0.22835   2.885 0.004134 **
sqrt(player20112015PPF100$X1stAssistGoal.xPer60)     0.27184    0.38790   0.701 0.483847
sqrt(player20112015PPF100$X2ndAssistGoal.xPer60)     0.38991    0.25002   1.560 0.119683
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.78 on 392 degrees of freedom
Multiple R-squared:  0.06747,   Adjusted R-squared:  0.04605
F-statistic: 3.151 on 9 and 392 DF,  p-value: 0.001102
```

2nd-Assists are the most significant variable. However, R-squared is 0.067 and it is very low. We probably should think about using the data at Equal-Handed situations as well.

Verification of Significant Variable



You can see a linear relationship between the 2nd-Assists and the next year's goals. However, this relationship is not as strong as the relationship at the Even-Handed situations.

I came up with a potential reason why the 2nd-Assists are statistically significant for power play situations. Most of the power play goals have assists. It is rare to see a team score Non-Assisted goals in the power play because you can have the puck and pass the puck in the offensive zone. Non-Assisted goals usually happen with turnovers. Turnovers usually do not happen at the power play because a short-handed team can throw the puck into the opponent's side without worrying about icing.
2nd-Assists mean the beginning of the offensive movement. Players who provide 2nd-Assists are the players who can carry the puck to the offensive zones or who can start the offensive movements. This is probably one of the reasons why 2nd-Assists are statistically significant in power play.

Set 2. Power Play and Equal-Handed Data in the previous year

As I mentioned, I have decided to use data at equal handed situation and power play situation because we do not have much data about power play situations.

Weighted Linear Model Result for 2011-2015 Season

Power Play situations are rarer than equal-handed situations.
The following is based on the data between 2011 and 2015 season for defensemen.
I used the power play and equal handed situation data in the following.
xPer60 is the power play situation and yPer60 is the equal-handed situation.

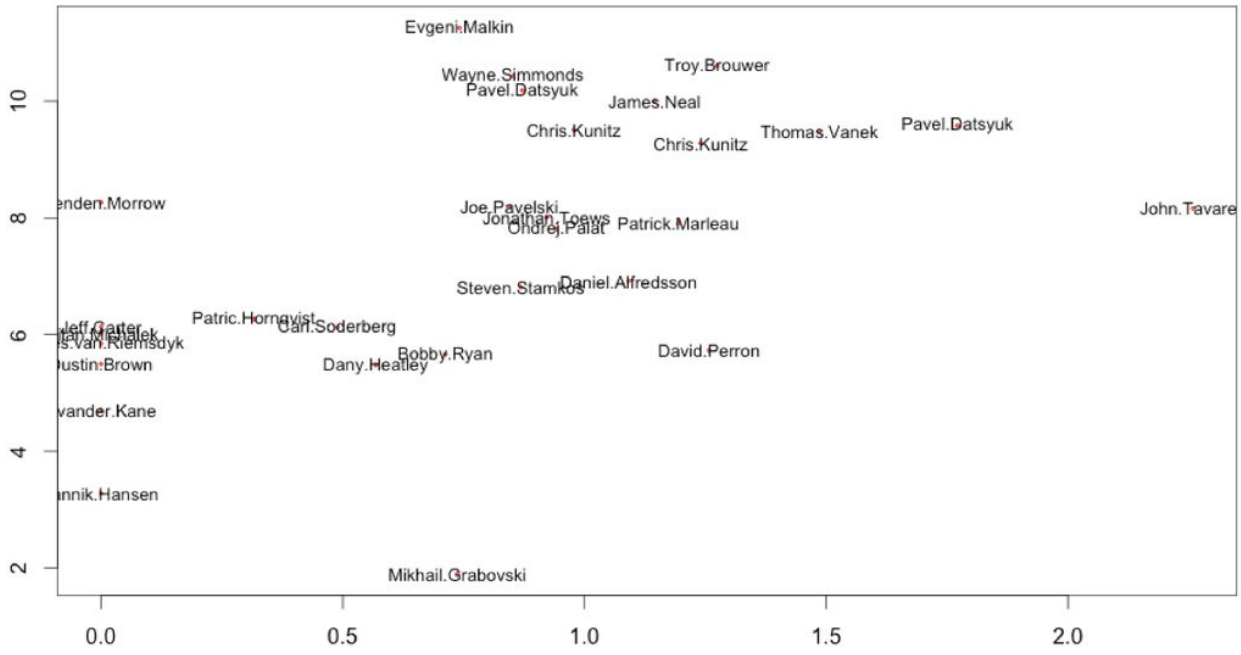
```
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       4.090756   0.726563   5.630  3.1e-08 ***
sqrt(player20112015PPEVF100$NonAssistGoal.xPer60) -0.866032   0.721966  -1.200 0.230921
sqrt(player20112015PPEVF100$X1stAssist.xPer60)     0.002676   0.347945   0.008 0.993867
sqrt(player20112015PPEVF100$X2ndAssisted1stAssist.xPer60)  0.060757   0.214162   0.284 0.776768
sqrt(player20112015PPEVF100$X2ndAssist.xPer60)     0.726710   0.213333   3.406 0.000715 ***
sqrt(player20112015PPEVF100$NonAssistPlayers.xPer60)  0.018204   0.409425   0.044 0.964554
sqrt(player20112015PPEVF100$X1stAssistPlayers.xPer60)  0.013085   0.251283   0.052 0.958491
sqrt(player20112015PPEVF100$X2ndAssistPlayers.xPer60)  0.639965   0.223192   2.867 0.004326 **
sqrt(player20112015PPEVF100$X1stAssistGoal.xPer60)  0.321551   0.368397   0.873 0.383199
sqrt(player20112015PPEVF100$X2ndAssistGoal.xPer60)  0.343391   0.219657   1.563 0.118656
sqrt(player20112015PPEVF100$NonAssistGoal.yPer60)  0.328088   0.569979   0.576 0.565152
sqrt(player20112015PPEVF100$X1stAssist.yPer60)     0.162666   0.526795   0.309 0.757623
sqrt(player20112015PPEVF100$X2ndAssisted1stAssist.yPer60)  0.286493   0.569542   0.503 0.615184
sqrt(player20112015PPEVF100$X2ndAssist.yPer60)     0.208778   0.584266   0.357 0.721003
sqrt(player20112015PPEVF100$NonAssistPlayers.yPer60)  0.720590   0.506839   1.422 0.155768
sqrt(player20112015PPEVF100$X1stAssistPlayers.yPer60)  -0.845696   0.546416  -1.548 0.122365
sqrt(player20112015PPEVF100$X2ndAssistPlayers.yPer60)  -0.055442   0.625336  -0.089 0.929391
sqrt(player20112015PPEVF100$X1stAssistGoal.yPer60)  -0.299634   0.541141  -0.554 0.580043
sqrt(player20112015PPEVF100$X2ndAssistGoal.yPer60)  0.846953   0.607998   1.393 0.164274
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.58 on 469 degrees of freedom
Multiple R-squared:  0.0786,    Adjusted R-squared:  0.04323
F-statistic: 2.223 on 18 and 469 DF,  p-value: 0.002815
```

2nd Assists and 2nd-Assisted players at the power play situation are two of the most significant variables. I come up with some potential reasons why they are significant.

1. At the power play situations, the most important thing is to keep the puck in the offensive zone and pass the puck to each other. 2nd-Assisted goals at the power play mean you can collaborate with other team members to score. Especially in power play, since you have a man advantage, you can have the puck in the opponent's zone. Thus, you need a player who can start passing the puck effectively. This is probably the reason why 2nd assists are important.

2. One of the potential reasons why 2nd-Assisted players are significant is dependence between players. If you have a chance to play with top forwards on the team in two consecutive seasons, even if you did not score nor assist, your team scores frequently. As a result, the team scores many goals when you are on the ice. For example, if a player played with Sidney Crosby at power-play situations, even just being on the ice with Sidney Crosby could help a team to score in the following year if he is still on the ice with Sidney Crosby.

Verification of Significant Variable



The relationship does not seem to be stronger than just using the power play data unfortunately. R-Squared improve a little and is now 0.0786. However, adjusted R-Squared is 0.04323.

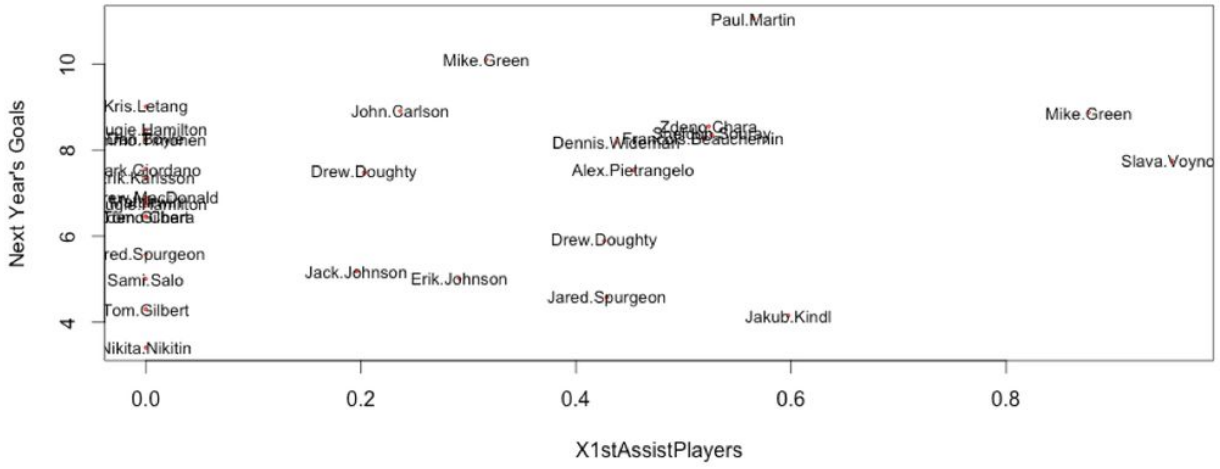
Power Play Situations (Defensemen)

Set 1. Only Power Play Data in the previous year

Weighted Linear Model Result for 2011-2015 Season

```
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        5.8254     0.7395   7.878 1.05e-12 ***
sqrt(player20112015PPD100$NonAssistGoal.xPer60)   -0.7931     0.9971  -0.795   0.4278
sqrt(player20112015PPD100$X1stAssist.xPer60)      -0.2509     0.6338  -0.396   0.6929
sqrt(player20112015PPD100$X2ndAssisted1stAssist.xPer60)  0.1692     0.3810   0.444   0.6578
sqrt(player20112015PPD100$X2ndAssist.xPer60)      -0.4490     0.3661  -1.226   0.2222
sqrt(player20112015PPD100$NonAssistPlayers.xPer60) -0.1355     0.5837  -0.232   0.8168
sqrt(player20112015PPD100$X1stAssistPlayers.xPer60)  0.9314     0.4135   2.253   0.0259 *
sqrt(player20112015PPD100$X2ndAssistPlayers.xPer60)  0.5022     0.3075   1.633   0.1048
sqrt(player20112015PPD100$X1stAssistGoal.xPer60)   0.1344     0.6118   0.220   0.8265
sqrt(player20112015PPD100$X2ndAssistGoal.xPer60)   0.2648     0.3503   0.756   0.4510
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.96 on 133 degrees of freedom
Multiple R-squared:  0.07655,   Adjusted R-squared:  0.01406
F-statistic: 1.225 on 9 and 133 DF,  p-value: 0.2849
```

Players on the ice when 1st-Assists are scored are the most significant variable. However, R-squared is 0.076 and it is very low. We probably think about using the data at Equal-Handed situations as well.



There seems to be a weak linear relationship between the 1st-Assist Players and the next year’s goals. However, this relationship is not as strong as the relationship at the Even-Handed situations unfortunately.

Set 2. Power Play and Equal-Handed Data in the previous year

As I mentioned, I have decided to use data at equal handed situation and power play situation because we do not have much data about power play situations.

Weighted Linear Model Result for 2011-2015 Season

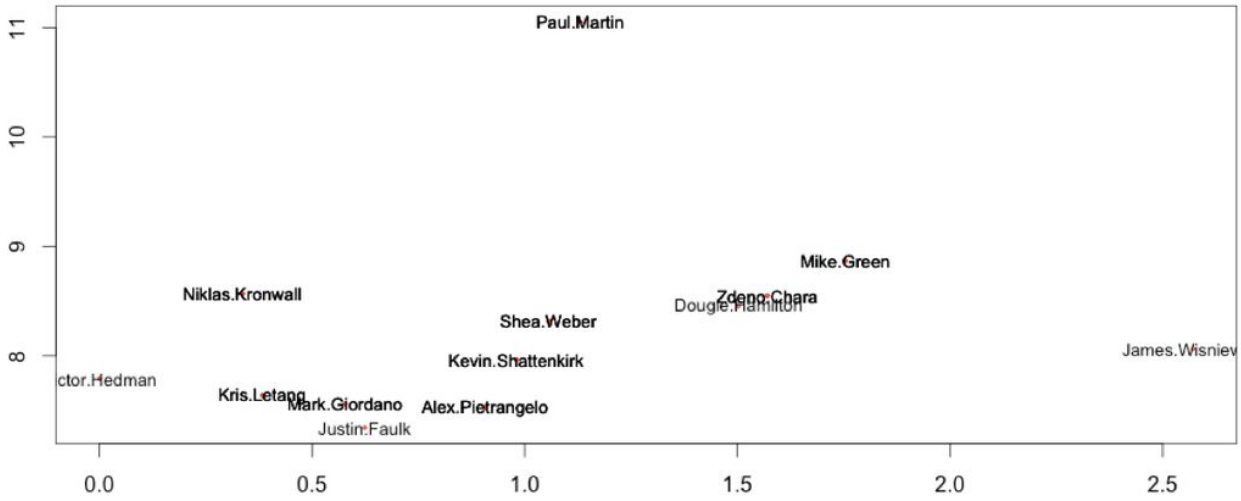
Power Play situations are rarer than equal-handed situations.
The following is based on the data between 2011 and 2015 season for defensemen.
I used both the power play situation and equal-handed situation data in the following.
xPer60 is the power play situation and yPer60 is the equal-handed situation.

```
Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                           4.1948     1.2776   3.283  0.00124 **
sqrt(player20112015PPEVD100$NonAssistGoal.xPer60)    -0.8880     1.1795  -0.753  0.45256
sqrt(player20112015PPEVD100$X1stAssist.xPer60)       -0.3037     0.6648  -0.457  0.64833
sqrt(player20112015PPEVD100$X2ndAssisted1stAssist.xPer60)  0.4141     0.3748   1.105  0.27079
sqrt(player20112015PPEVD100$X2ndAssist.xPer60)       -0.2741     0.3874  -0.708  0.48012
sqrt(player20112015PPEVD100$NonAssistPlayers.xPer60) -0.3221     0.6393  -0.504  0.61499
sqrt(player20112015PPEVD100$X1stAssistPlayers.xPer60)  0.3597     0.4220   0.852  0.39527
sqrt(player20112015PPEVD100$X2ndAssistPlayers.xPer60) -0.1411     0.3500  -0.403  0.68737
sqrt(player20112015PPEVD100$X1stAssistGoal.xPer60)    0.5371     0.7285   0.737  0.46198
sqrt(player20112015PPEVD100$X2ndAssistGoal.xPer60)    0.9299     0.3639   2.555  0.01148 *
sqrt(player20112015PPEVD100$NonAssistGoal.yPer60)     1.0957     1.4406   0.761  0.44797
sqrt(player20112015PPEVD100$X1stAssist.yPer60)        0.5242     0.9840   0.533  0.59493
sqrt(player20112015PPEVD100$X2ndAssisted1stAssist.yPer60) -0.3426     1.0215  -0.335  0.73773
sqrt(player20112015PPEVD100$X2ndAssist.yPer60)        0.8722     1.1394   0.766  0.44500
sqrt(player20112015PPEVD100$NonAssistPlayers.yPer60) -1.8157     0.9479  -1.916  0.05710 .
sqrt(player20112015PPEVD100$X1stAssistPlayers.yPer60)  1.9416     1.1591   1.675  0.09575 .
sqrt(player20112015PPEVD100$X2ndAssistPlayers.yPer60)  0.4709     1.0563   0.446  0.65630
sqrt(player20112015PPEVD100$X1stAssistGoal.yPer60)    1.3086     1.1256   1.163  0.24663
sqrt(player20112015PPEVD100$X2ndAssistGoal.yPer60)   -1.3645     1.0436  -1.308  0.19280
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68.73 on 171 degrees of freedom
Multiple R-squared:  0.1081,    Adjusted R-squared:  0.01419
F-statistic: 1.151 on 18 and 171 DF,  p-value: 0.3078
```

A 2nd-Assisted goal at power play is the most significant variable in the power play situations. This suggests that even at power play situations, teamwork is important. At power play situations, defensemen have a good chance to take a shot. Thus the ability to put a puck into the net is important.

However, I need to think about the limitation of this model because there are only 190 players. There are only 190 defensemen in four seasons who play constantly at power play situations because at power play situations, a team only uses one defenseman or at most two defensemen.

Verification of Significant Variable



There are some good defencemen on the chart such as Shea Weber, Alex Pietrangelo and Zdeno Chara. Alex Pietrangelo, Shea Weber and Zdeno Chara are all on the power play list. You can see the linear relationship between 1st-Assisted goal and next year's team goals.

VI. Analysis with Different Variables (Cross-Validation)

In **V. Analysis of Linear Models** I used a weighted multiple linear regression to predict how many goals a team can score while a player is on the ice per 60 minutes because I assume a linear relationship between explanatory variables per 60 minutes and the number of goals a team can score while a player is on the ice per 60 minutes.

In **VI. Analysis with Different Variables**, I have decided to use five different sets of explanatory variables for the linear model. Then I used to cross-validation to figure out the best set of explanatory variables.

Set 1: All the Varaibles*
Set 2: Remove variables whose p-values are bigger than 0.05 one by one from Set 1
Set 3: Remove variables whose p-values are bigger than 0.1 one by one from Set 1
Set 4: Use Step function to include interaction terms of Set 2 in both directions
Set 5: Use Step function to include interaction terms of Set 3 in both directions

All the variables are the ones with red ink in the following table.

Variable Type	1st Assisted Goals	2nd Assisted Goals	Non-assisted Goals
Goal	1stAssistGoal	2ndAssistGoal	NonAssistGoal
Assist	1stAssist	2ndAssisted1stAssist	
		2ndAssist	
Players on The Ice	1stAssistPlayers	2ndAssistPlayers	NonAssistPlayers

First I look at the data of forwards at equal-handed situation.

Equal Handed Situations (Forwards)

There are about 250 players who play more than 500 minutes in each two consecutive seasons.
I have data of four seasons from 2012 to 2015. Thus, there are three sets of years to compare (2012-2013, 2013-2014 and 2014-2015).
As a result, I have data of 766 players.

I ran the following method 10000 times.
Each time I divide 766 players into a train set (600 players) and a test set (166 players). Based on the training set, I made a linear model and fit it to the test set. I calculated the sum of squared errors of prediction (sse) and chose the best one which has the smallest average sum of squared errors of prediction.

Set of Explanatory Variables

Set 1: All the Variables
Set 2:

1. NonAssistGoal
2. 2ndAssisted1stAssist
3. 2ndAssist
4. 2ndAssistGoal

Set 3:

1. NonAssistGoal
2. 2ndAssisted1stAssist
3. 2ndAssist
4. 2ndAssistGoal
5. 1stAssist
6. 2ndAssistPlayers

Set 4:

1. NonAssistGoal
2. 2ndAssisted1stAssist
3. 2ndAssist
4. 2ndAssistGoal
5. 2ndAssisted1stAssist:2ndAssist

Set 5:

1. NonAssistGoal
2. 2ndAssisted1stAssist

3. 2ndAssist
4. 2ndAssistPlayers
5. 2ndAssistGoal
6. 2ndAssisted1stAssist:2ndAssist
7. NonAssistGoal:2ndAssist
8. 2ndAssisted1stAssist:2ndAssistPlayers

The following table is a summary of SSEs for each set.

	Set 1	Set 2	Set 3	Set 4	Set 5
SSE(sum of squared errors)	0.429	0.429	0.428	0.421	0.417

I decided to choose set 5 as the best set of explanatory variables because it has the smallest SSE. The following is the table of coefficients for the linear model with Set 5.

```
Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                           1.6352     0.3757   4.353 1.59e-05 ***
sqrt(NonAssistGoal.xPer60)                            1.2603     0.6227   2.024   0.0434 *
sqrt(X2ndAssisted1stAssist.xPer60)                   -0.6757     0.5990  -1.128   0.2597
sqrt(X2ndAssist.xPer60)                              -0.1357     0.4858  -0.279   0.7801
sqrt(X2ndAssistPlayers.xPer60)                       -0.5359     0.5611  -0.955   0.3399
sqrt(X2ndAssistGoal.xPer60)                           0.9681     0.1445   6.698 4.93e-11 ***
sqrt(X2ndAssisted1stAssist.xPer60):sqrt(X2ndAssist.xPer60)  1.1658     0.6950   1.678   0.0940 .
sqrt(NonAssistGoal.xPer60):sqrt(X2ndAssist.xPer60)   -1.4399     0.9552  -1.507   0.1322
sqrt(X2ndAssisted1stAssist.xPer60):sqrt(X2ndAssistPlayers.xPer60)  1.1671     0.8145   1.433   0.1524
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.91 on 591 degrees of freedom
Multiple R-squared:  0.2126,    Adjusted R-squared:  0.2019
F-statistic: 19.95 on 8 and 591 DF,  p-value: < 2.2e-16
```

In this model, 2nd-Assisted Goal is the most important variable to predict the number of goals. 2nd-Assisted Goal is the team play goal and it makes sense that this is an important variable to predict the number of team goals in the following season because 2nd-Assisted Goals mean the player can collaborate with other players and score goals. Also 2nd-Assisted Goals mean you have the ability to shoot the puck accurately.

In the above interaction model, there are some explanatory variables which do not seem significant. Also it is hard to interpret the result if you have many explanatory variables. I have decided to remove an insignificant interaction term one by one, which means its p-value is bigger than 0.05. I removed the following two terms.

1. sqrt(NonAssistGoal.xPer60):sqrt(X2ndAssist.xPer60)

2. sqrt(X2ndAssisted1stAssist.xPer60):sqrt(X2ndAssistPlayers.xPer60)

```
Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                              1.35522    0.24907   5.441 7.15e-08 ***
sqrt(NonAssistGoal.xPer60)                               0.34460    0.14553   2.368   0.0181 *
sqrt(X2ndAssisted1stAssist.xPer60)                       0.05318    0.38663   0.138   0.8906
sqrt(X2ndAssist.xPer60)                                 -0.47971    0.40617  -1.181   0.2379
sqrt(X2ndAssistPlayers.xPer60)                           0.27534    0.13845   1.989   0.0471 *
sqrt(X2ndAssistGoal.xPer60)                              0.78208    0.12951   6.039 2.43e-09 ***
sqrt(X2ndAssisted1stAssist.xPer60):sqrt(X2ndAssist.xPer60)  1.43714    0.59762   2.405   0.0164 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.95 on 759 degrees of freedom
Multiple R-squared:  0.2017,    Adjusted R-squared:  0.1954
F-statistic: 31.96 on 6 and 759 DF,  p-value: < 2.2e-16
```

This model shows the importance of 2nd-Assisted Goals. 2nd-Assisted Goals itself is an important explanatory variable. The interaction term of 2nd-Assisted Goals' 1st Assists and 2nd-Assists is also an important variable. This means that the effect of 2nd-Assisted 1st Assists becomes larger as 2nd-Assists gets larger.

This model makes sense because 2nd Assists are the result of team play. Also as I mentioned in the **V. Analysis of Linear Models,** there are some players who are better at scoring goals than providing assists such as Alex Ovechkin and Steven Stamkos. Thus I guess NonAssistedGoals are statistically significant because of those players.

Predicted Goals in 2015-2016

The following is the table of predicted team goals for forwards at even handed situation in 2015-2016 season. Almost all the star plays are on this list such as Sidney Crosby, Ryan Getzlaf and Patrick Kane. Some variables are correlated such as a pair of Corey Perry and Ryan Getzlaf and a pair of Jamie Benn and Tyler Seguin. They were on the first line of Anaheim Ducks and Dallas Stars last year respectively.

Unfortunately, Anaheim Ducks has scored only 56 goals in 30 games this year and its offence has the worst number of goals in the league. Corey Perry and Ryan Getzlaf have had a total of only 37 points in 30 games so far. Anaheim Ducks has finished the division with the first place for the past three consecutive seasons. This season is exceptionally bad for them.

On the contrary, Jamie Benn and Tyler Seguin have had a total of 83 points in 32 games so far. Dallas as a team has scored 108 goals in 32 games. They are producing more goals than I predict here.

Name	GoalPredict
Vladimir.Tarasenko	2.838228
Blake.Wheeler	2.843286
Evgeni.Malkin	2.843879
Jakub.Voracek	2.848884
Chris.Kunitz	2.856241
Logan.Couture	2.860211
Jason.Spezza	2.860575
Claude.Giroux	2.861811
James.Neal	2.885846
Joe.Thornton	2.889230
Patrick.Sharp	2.899544
John.Tavares	2.903942
Taylor.Hall	2.904766
Jonathan.Toews	2.905067
Gabriel.Landeskog	2.905379
Paul.Stastny	2.908333
Mathieu.Perreault	2.909491
Kyle.Okposo	2.918610
Patrick.Kane	2.922211
Sidney.Crosby	2.927517
Matt.Duchene	2.993856
Henrik.Zetterberg	3.004026
Jamie.Benn	3.041383
Tyler.Seguin	3.044260
Steven.Stamkos	3.048507
Gustav.Nyquist	3.048602
Corey.Perry	3.131814
Ryan.Getzlaf	3.151162

First I look at the data of defensemen at equal-handed situation.

Equal Handed Situations (Defensemen)

There are about 120 players who play more than 500 minutes in each two consecutive seasons. I have data of four seasons from 2012 to 2015. Thus, there are three sets of years to compare (2012-2013, 2013-2014,2014-2015).
As a result, I have data of 364 players.

I ran the following method 10000 times.
I decided to divide 364 players into a train set (300 players) and a test set (64 players). Based on the training set, I made a linear model and fit it to the test set. I calculated the sum of squared errors of prediction (sse) and chose the best one which has the smallest average sum of squared errors of prediction.

Set 2 & Set 3 have the same variables. Thus Set 4 & Set 5 have the same variables.

Set of Explanatory Variables

Set 1: All the Variabless
Set 2 & Set 3:

1. 1stAssist
2. 2ndAssisted1stAssist
3. 2ndAssistPlayers
4. 2ndAssistGoal

Set 4 & Set 5:

1. 1stAssist
2. 2ndAssisted1stAssist
3. 2ndAssistPlayers
4. 2ndAssistGoal
5. 2ndAssisted1stAssist:2ndAssistGoal

	Set 1	Set 2 & 3	Set 4 & 5
SSE(sum of squared errors)	0.2688	0.265	0.2556

I decided to choose Set 4 & Set 5 as the best set of explanatory variables because they have the minimum SSE.

```
Coefficients:
                                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                1.9168     0.1931   9.925   <2e-16 ***
sqrt(X1stAssist.xPer60)                                    0.3632     0.1586   2.290   0.0226 *
sqrt(X2ndAssisted1stAssist.xPer60)                        -0.1948     0.3072  -0.634   0.5264
sqrt(X2ndAssistPlayers.xPer60)                             0.3333     0.1548   2.153   0.0320 *
sqrt(X2ndAssistGoal.xPer60)                               -0.1450     0.3627  -0.400   0.6896
sqrt(X2ndAssisted1stAssist.xPer60):sqrt(X2ndAssistGoal.xPer60)   1.6841     0.7999   2.105   0.0360 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.61 on 358 degrees of freedom
Multiple R-squared:  0.1193,    Adjusted R-squared:  0.107
F-statistic: 9.699 on 5 and 358 DF,  p-value: 1.06e-08
```

The 2nd-Assisted Goal’s 1st Assist and 2nd-Assisted Goal are not significant by themselves. But their interaction term is significant. Thus, I decided not to remove any variables.
It makes sense that variables associated with 2nd-Assisted goals are significant because when a defenseman usually scores or assists, a team is usually at the offensive zone and passes the puck. When a team passes the puck at the offensive zone, usually 2nd-Assisted goals or 1st-Assisted goals are scored. I think this is also one of the reasons why 1st Assists are important.

It is interesting to know that players on the ice when a 2nd-Assisted goal is scored are important. One of the possible reason is that it is important to have defencemen who can keep the puck in the offensive zone for a team to score. Another possible reason is that if you play with a very good forward in the previous year and also play with him this year, your team also has a good chance to score this year. For example, if you played with Steven Stamkos last year and also played with him this year, your team would have a better chance to score even if you did not contribute to a goal directly.

Predicted Goals in 2015-2016

The table on the next page is predicted team goals per 60 minutes for defensemen at even handed situation in 2015-2016 season. The number of predicted goals for top defenders is smaller than the number of predicted goals for top forwards. This is understandable because the role of defensemen is to defend goals rather than to score goals.

Almost all the offensive defenders are on this list such as Erik Karlsson, Alex Pietrangelo and Victor Hedman are on the list.

I list two interesting discoveries in this table.

1) Corey Perry and Ryan Getzlaf are two of the top forwards in terms of team scoring at even handed situation as you can see in the above. Hampus Lindholm and Sami Vatanen are both defencemen in Anaheim Ducks. I guess Hampus Lindholm and Sami Vatanen had a chance to play with a pair of Corey Perry and Ryan Getzlaf and the team scored many goals while they were on the ice.That is probably the reason why the predicted goals of Hampus Lindholm and Sami Vatanen are high.
2) St. Louis Blues have two defensemen, Jay Bouwmeester and Alex Pietrangelo on this list. They play together as the first defense line of St. Louis Blues. Alex Pietrangelo had 46 points, but Jay Bouwmeester had only 13 points in 2014-2015 season. I came up with two potential reasons why a team with Jay Bouwmeester is expected to score many goals.
    a) The first reason is that St. Louis Blues indeed had very good forwards such as Vladimir Tarasenko, David Backes, T.J. Oshie and Alex Steen in 2014-2015 season. These four players had 250 points total. As a result, the forwards scored many goals while Jay Bouwmeester and Alex Pietrangelo were on the ice.
    b) Another possible reason is that Jay Bouwmeester may be so good at defense that other players can focus on offense. This second reason shows the limitation of this model. In ice hockey, defence and offence are highly dependent. Even if you are exceptionally good at offense, your team cannot score many goals if forwards on the team have to pay considerable amount of attention to defense. If you have a very good pair of defensemen, forwards on the team can focus on scoring. As a result, the team as a whole can increase the number of goals.

Name	GoalPredict
Justin.Schultz	2.595469
Erik.Johnson	2.596625
Jared.Spurgeon	2.624853
Alex.Goligoski	2.631597
Trevor.Daley	2.644081
Jacob.Trouba	2.648017
Ryan.Ellis	2.649586
Brent.Seabrook	2.655429
Dustin.Byfuglien	2.655652
Torey.Krug	2.659099
Dougie.Hamilton	2.659314
Erik.Karlsson	2.659828
Jay.Bouwmeester	2.681810
Mark.Giordano	2.683737
Alex.Pietrangelo	2.701894
Victor.Hedman	2.727026
Sami.Vatanen	2.733694
Hampus.Lindholm	2.763553
Tyson.Barrie	2.928306

Power Play Situations (Forwards)

I have decided to use both power play and equal-handed situation data to make a model for power play situations because the information at only power play situation is limited.
There are about 160 players who play more than 100 minutes in each two consecutive seasons. I have data of four seasons from 2012 to 2015. Thus, there are three sets of years to compare (2012-2013, 2013-2014 and 2014-2015).
As a result, I have data of 488 players.

I ran the following method 10000 times.
Each time I divide 488 players into a train set (400 players) and a test set (88 players). Based on the training set, I made a linear model and fit it to the test set. I calculated the sum of squared errors of prediction (sse) and chose the best one which has the smallest average sum of squared errors of prediction.

Set of Explanatory Variables

Set 1: All the Variables

Set 2:

1. 2ndAssist.xPer60
2. 2ndAssistPlayers.xPer60
3. 2ndAssistGoal.xPer60
4. 2ndAssistGoal.yPer60

Set 3:

1. 2ndAssist.xPer60
2. 2ndAssistPlayers.xPer60
3. 2ndAssistGoal.yPer60

Set 4:

1. 2ndAssist.xPer60
2. 2ndAssistPlayers.xPer60
3. 2ndAssistGoal.yPer60
4. 2ndAssist.xPer60:2ndAssistGoal.yPer60

Set 5:

1. 2ndAssist.xPer60
2. 2ndAssistPlayers.xPer60
3. 2ndAssistGoal.xPer60
4. 2ndAssistGoal.yPer60

5. 2ndAssistGoal.xPer60:2ndAssistGoal.yPer60

Table of SSE

	Set 1	Set 2	Set 3	Set 4	Set 5
SSE(sum of squared errors)	0.503	0.479	0.480	0.467	0.463

I decide to choose Set 5 as the best set of explanatory variables.
The following is the table of coefficients of this model for the entire data set.

xPer60 is the power play situation and yPer60 is the equal-handed situation.

```
Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                         5.5723     0.9402   5.927 5.89e-09 ***
sqrt(X2ndAssist.xPer60)                             0.8653     0.2101   4.118 4.50e-05 ***
sqrt(X2ndAssistPlayers.xPer60)                      0.6030     0.2300   2.622  0.00902 **
sqrt(X2ndAssistGoal.xPer60)                        -1.5833     0.8557  -1.850  0.06490 .
sqrt(X2ndAssistGoal.yPer60)                        -1.2414     1.2784  -0.971  0.33201
sqrt(X2ndAssistGoal.xPer60):sqrt(X2ndAssistGoal.yPer60)   2.5538     1.1057   2.310  0.02133 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65.73 on 482 degrees of freedom
Multiple R-squared:  0.0739,    Adjusted R-squared:  0.0643
F-statistic: 7.693 on 5 and 482 DF,  p-value: 5.667e-07
```

This model shows the importance of 2nd-Assisted Goals in power play situations. 2nd-Assisted Goals, 2nd Assists, players on the ice when a 2nd-assisted goal is scored are significant.
2nd Assists and players on the ice when a 2nd-assisted goal is scored are significant by themselves.
The importance of 2nd Assists is understandable because 2nd Assists mean you can collaborate with other players to score goals at power play. In power play situations, you need players who can pass the puck with other teammates in the offensive zone because your team has a man advantage and you can possess the puck.

A possible reason why players on the ice when a 2nd-Assisted goal is scored is important is that if you played with a very good forward in the previous year and also play with him this year, your team has a better chance to score this year. For example, if you played with Steven Stamkos last year and also played with him this year, your team has a better chance to score even if you did not contribute to a goal directly.

2nd-Assisted Goals at both power plays is magnified if you score 2nd-Assisted Goals at even-handed situations.
These analyses suggest that even at power play situation, the most important thing is teamwork because 2nd-Assisted Goals are a result of teamwork.

Predicted Goals in 2015-2016

The following is the table of predicted team goals per 60 minutes for forwards at power play situation in 2015-2016 season based on the best predictors.

Some teams have several members on this list such as Pittsburgh Penguins, Chicago Blackhawks and Boston Bruins. Pittsburgh Penguins has Chris Kunitz and Sidney Crosby. Boston Bruins has Milan Lucic, Loui Eriksson and Patrice Bergeron. Chicago Blackhawks has Patrick Kane and Marian Hossa.

This makes sense because only good offensive players can play in the power play situation. Thus players on the team are highly correlated.

Name	GoalPredict
Ryan.Johansen	7.482072
Patrick.Kane	7.489013
Milan.Lucic	7.492835
Nicklas.Backstrom	7.494789
Gustav.Nyquist	7.502582
James.Neal	7.512527
Loui.Eriksson	7.516435
Patrice.Bergeron	7.529596
Claude.Giroux	7.537237
Marian.Hossa	7.540345
Jakub.Voracek	7.550535
Troy.Brouwer	7.567365
John.Tavares	7.571438
Joe.Pavelski	7.572858
Patrik.Elias	7.603005
Sidney.Crosby	7.641657
Nick.Bonino	7.643925
Corey.Perry	7.690707
Chris.Kunitz	7.878308
Thomas.Vanek	7.908821
Steven.Stamkos	8.129572

Power Play Situations (Defensemen)

I have decided to use both power play and equal-handed situation data to make a model for power play situations for defencemen as well because the information at only power play situation is limited..
There are about 60 players who play more than 50 minutes in each two consecutive seasons.
I have data of four seasons from 2012 to 2015. Thus, there are three sets of years to compare (2012-2013, 2013-2014 and 2014-2015).
As a result, I have data of 190 players.
xPer60 is the power play situation and yPer60 is the equal-handed situation.

I ran the following method 10000 times.
Each time I divide 190 players into a train set (160 players) and a test set (30 players). Based on the training set, I made a linear model and fit it to the test set. I calculated the sum of squared errors of prediction (sse) and chose the best one which has the smallest average sum of squared errors of prediction.

Set of Explanatory Variables

Set 1: All the Variables

Set 2: (P-Value<0.05)

1. 2ndAssistGoal.xPer60
2. NonAssistPlayers.yPer60,

Set 3: (P-Value<0.1)

1. 2ndAssistGoal.xPer60
2. NonAssistPlayers.yPer60
3. 1stAssistPlayers.yPer60

Set 4: (Stepwise of Set 2)

1. 2ndAssistGoal.xPer60

Set 5: (Stepwise of Set 3)

1. 2ndAssistGoal.xPer60

Table of SSE

	Set 1	Set 2	Set 3	Set 4	Set 5
SSE(sum of squared errors)	1.37	1.28	1.29	1.36	1.36

I decide to choose set 2 as the best set of explanatory variables because it has the smallest SSE. The following is the table of coefficients for this model.

xPer60 is the power play situation and yPer60 is the equal-handed situation.

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.8228     0.7660   6.296 2.15e-09 ***
sqrt(X1stAssistPlayers.yPer60)  1.8266     1.0908   1.675  0.09570 .
sqrt(X2ndAssistGoal.xPer60)     0.9225     0.3343   2.760  0.00637 **
sqrt(NonAssistPlayers.yPer60)  -1.7198     0.8990  -1.913  0.05729 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67.39 on 186 degrees of freedom
Multiple R-squared:  0.06737,   Adjusted R-squared:  0.05232
F-statistic: 4.478 on 3 and 186 DF,  p-value: 0.004622
```

The significance of 2nd-Assisted Goal is comprehensible because especially for defensemen, the ability to score when your team possess the puck is important at power play situations. When your team possesses the puck, usually your team scores goals with 2nd-Assists.

It is understandable that the coefficient of Non-Assisted Players is negative because this means that a player did not contribute to a non-assisted goal when a non-assisted goal is scored at equal-handed situation. Thus, he probably did not have strong ability to contribute to the team offence.

The significance of 1stAssistPlayers is questionable. It does not make much sense the coefficient of 1stAssistPlayers at equal-handed situation is positive. I tried to come up with possible solutions.
One possible explanation is the dependence between players. Being on the ice in 2014 with Sidney Crosby could help with your goal scoring in 2015 if you are still on the ice with Sidney Crosby. However, this explanation is insufficient because 1stAssistPlayers at equal-handed situation is significant when you predict the number of goals a team can score at power plays. Equal-handed situations and power play situations are different.

The p-value of 1stAssistPlayers at equal-handed situation is 0.095, which is almost 0.1. This means that some researchers think about removing this variable. I think it makes sense to remove this because it is hard to understand the importance of this variable.

The following is the result after I removed this variable.

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.9115     0.4071  14.521  < 2e-16 ***
sqrt(X2ndAssistGoal.xPer60)    0.9578     0.3352   2.857  0.00476 **
sqrt(NonAssistPlayers.yPer60) -1.5410     0.8970  -1.718  0.08744 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67.71 on 187 degrees of freedom
Multiple R-squared:  0.05331,   Adjusted R-squared:  0.04318
F-statistic: 5.265 on 2 and 187 DF,  p-value: 0.005965
```

Still the coefficient of Non-Assisted Players is negative. However, its p-values is 0.087 and its statistical significance is a bit questionable. 2nd-Assisted Goal seems to be a very important explanatory variable.

Predicted Goals in 2015-2016

The following is the table of predicted team goals per 60 minutes for defensemen at power play situation in 2015-2016 season based on the best predictors.

Only two teams have several members on this list. Minnesota Wild and Boston Bruins.

Minnesota Wild has Jonas Brodin and Jared Spurgeon.

This makes sense because most of the teams only use one defensemen and at most two. Thus, defensemen on each team is not as correlated as forwards.

Name	GoalPredict
Mike.Green	6.373646
Alec.Martinez	6.452058
Alexander.Edler	6.519441
Dan.Boyle	6.535511
Kevin.Shattenkirk	6.537760
Jonas.Brodin	6.566632
Torey.Krug	6.591430
Mark.Streit	6.604726
Mark.Giordano	6.674225
Jared.Spurgeon	6.798439
Dennis.Wideman	6.807578
Dustin.Byfuglien	6.874618
Sami.Vatanen	6.889264
Tyson.Barrie	6.895710
Zdeno.Chara	6.972542

VII. Limitations

Independence

This study examines the offensive contribution of each player. Since hockey is a team sports, almost all the data is strongly correlated. Thus, it is hard to pinpoint the ability of each player. When I wrote the analysis in section ***VI. Analysis with Different Variables***, I mention the dependence between players several times. It is hard to evaluate the true ability of individual players.

Defense

In this analysis, I examine the ability to contribute to team goals based on goals and assists a team scored for. However, the offense of team is deeply correlated with the defense of the team. A team with good defense is able to produce more goals because forwards can focus on the offense. In order to understand both defense and offence, you have to take goals against into consideration.

VIII. Conclusions

The above data analysis reveals the importance of teamwork to score goals. You should acquire a player who can work with other players and score rather than acquire a player that scores a non-assisted goal.
The following the summary of each situation and position.

Equal-Handed (Forwards)

If you want to acquire a forward that can increase the number of team goals at equal-handed situations, you should look at the stats of Non-Assisted Goals and 2nd-Assisted Goals.
Non-Assisted Goals are a good measure of the individual ability to score. 2nd-Assisted Goals are the ability to produce goals with other players. If you have a choice to acquire a player, you should acquire a player who has better stats at 2nd-Assisted Goals than Non-Assisted Goals because coefficients of 2nd-Assisted Goals' variables are bigger than those of Non-Assisted Goals' variables.

What you should look at
Statistics of 2nd-Assisted Goals & Non-Assisted Goals

Result of Linear Regression

```
Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                              1.35522    0.24907   5.441 7.15e-08 ***
sqrt(NonAssistGoal.xPer60)                               0.34460    0.14553   2.368   0.0181 *
sqrt(X2ndAssisted1stAssist.xPer60)                       0.05318    0.38663   0.138   0.8906
sqrt(X2ndAssist.xPer60)                                 -0.47971    0.40617  -1.181   0.2379
sqrt(X2ndAssistPlayers.xPer60)                           0.27534    0.13845   1.989   0.0471 *
sqrt(X2ndAssistGoal.xPer60)                              0.78208    0.12951   6.039 2.43e-09 ***
sqrt(X2ndAssisted1stAssist.xPer60):sqrt(X2ndAssist.xPer60)  1.43714    0.59762   2.405   0.0164 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.95 on 759 degrees of freedom
Multiple R-squared:  0.2017,    Adjusted R-squared:  0.1954
F-statistic: 31.96 on 6 and 759 DF,  p-value: < 2.2e-16
```

Equal-Handed (Defensemen)

If you want to acquire a defenseman that can increase the number of team goals at equal-handed situations, you should look at the statistics of 1st Assists and 2nd-Assisted Goals.
It is rare for a defenseman to score non-assisted goals. Thus, you do not have to pay much attention to the statistics of non-assisted goals.
You should look at the statistics of 1st assists and variables related to 2nd-Assisted goals If you have a choice to acquire a player, you should acquire a player who has better stats at 2nd-Assisted Goals than 1st-Assisted Goals because coefficients of 2nd-Assisted Goals' variables are bigger than those of 1st-Assisted Goals' variables.

What you should look at

Statistics of 2nd-Assisted Goals & 1st-Assisted Goals

Result of Linear Regression

```
Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                               1.9168     0.1931   9.925   <2e-16 ***
sqrt(X1stAssist.xPer60)                                   0.3632     0.1586   2.290   0.0226 *
sqrt(X2ndAssisted1stAssist.xPer60)                       -0.1948     0.3072  -0.634   0.5264
sqrt(X2ndAssistPlayers.xPer60)                            0.3333     0.1548   2.153   0.0320 *
sqrt(X2ndAssistGoal.xPer60)                              -0.1450     0.3627  -0.400   0.6896
sqrt(X2ndAssisted1stAssist.xPer60):sqrt(X2ndAssistGoal.xPer60)   1.6841     0.7999   2.105   0.0360 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.61 on 358 degrees of freedom
Multiple R-squared:  0.1193,    Adjusted R-squared:  0.107
F-statistic: 9.699 on 5 and 358 DF,  p-value: 1.06e-08
```

Power-Play (Forwards)

If you want to acquire a forward that can increase the number of team goals at power-play situations, you should look at the statistics of 2nd-Assisted Goals.
2nd-Assisted Goals represent the ability to produce goals with other players. At power-play situations, the ability to score a goal by teamwork is essential.
If a forward scores many goals not only at even-handed situations but also at power play situations, that would be a huge plus.

What you should look at

Statistics of 2nd-Assisted Goals at both Even Handed and Power Play

Result of Linear Regression

```
Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                         5.5723     0.9402   5.927 5.89e-09 ***
sqrt(X2ndAssist.xPer60)                             0.8653     0.2101   4.118 4.50e-05 ***
sqrt(X2ndAssistPlayers.xPer60)                      0.6030     0.2300   2.622  0.00902 **
sqrt(X2ndAssistGoal.xPer60)                        -1.5833     0.8557  -1.850  0.06490 .
sqrt(X2ndAssistGoal.yPer60)                        -1.2414     1.2784  -0.971  0.33201
sqrt(X2ndAssistGoal.xPer60):sqrt(X2ndAssistGoal.yPer60)  2.5538     1.1057   2.310  0.02133 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65.73 on 482 degrees of freedom
Multiple R-squared:  0.0739,    Adjusted R-squared:  0.0643
F-statistic: 7.693 on 5 and 482 DF,  p-value: 5.667e-07
```

Power-Play (Defensemen)

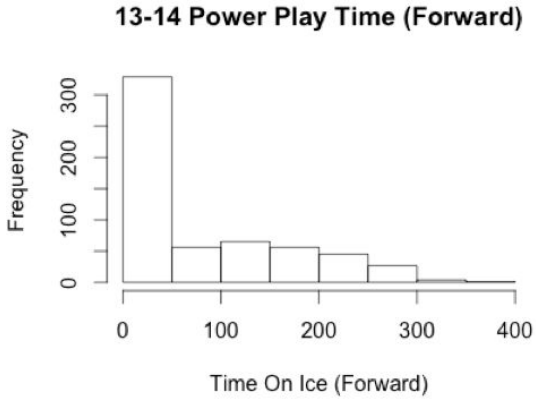
If you want to acquire a defenseman that can increase the number of team goals at power-play situations, you should look at the statistics of 2nd-Assisted Goals.
2nd-Assisted Goals represent the ability to produce goals with other players. Also this represents the ability to shoot and put the puck into the net.
You should not acquire a defenseman who is just on the ice when the non-assisted goal is scored because it has negative coefficient. This probably means that this defenseman does not have strong ability to contribute to the team offense.

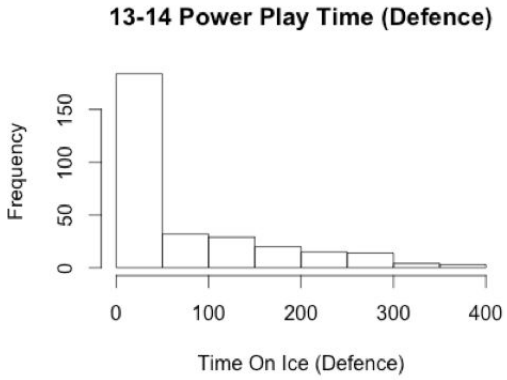
What you should look at
Statistics of 2nd-Assisted Goals at Even Handed & Non-Assisted Goals

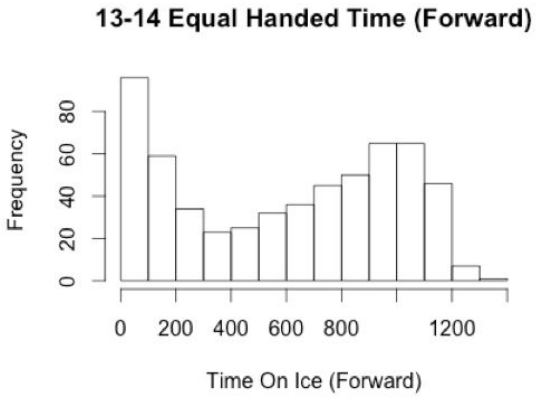
Result of Linear Regression

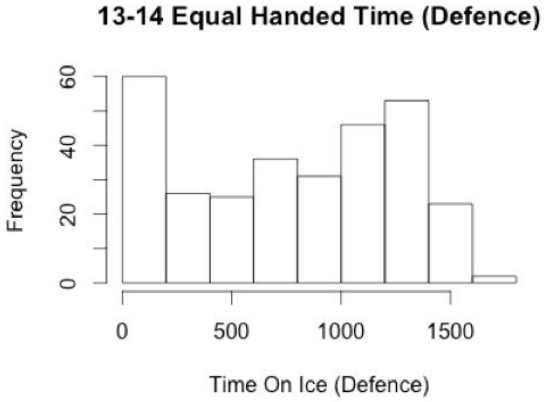
```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.9115     0.4071  14.521  < 2e-16 ***
sqrt(X2ndAssistGoal.xPer60)     0.9578     0.3352   2.857  0.00476 **
sqrt(NonAssistPlayers.yPer60)  -1.5410     0.8970  -1.718  0.08744 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67.71 on 187 degrees of freedom
Multiple R-squared:  0.05331,   Adjusted R-squared:  0.04318
F-statistic: 5.265 on 2 and 187 DF,  p-value: 0.005965
```

Appendix A: Summary of Time on Ice
2013-2014 Season

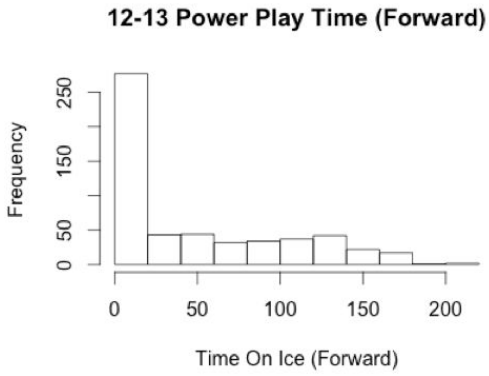


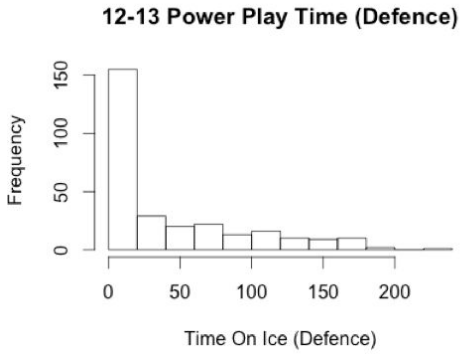


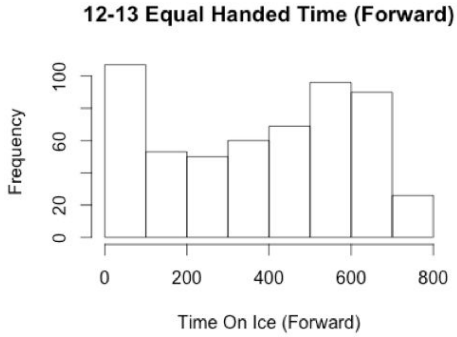


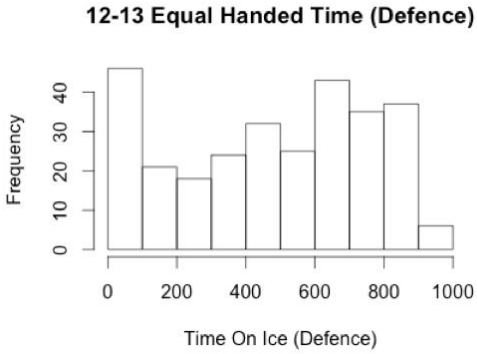


2012-2013 Season



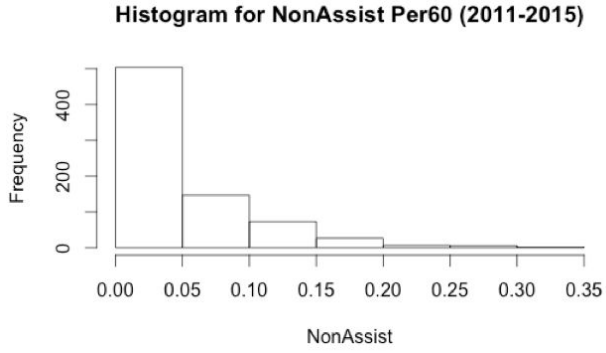


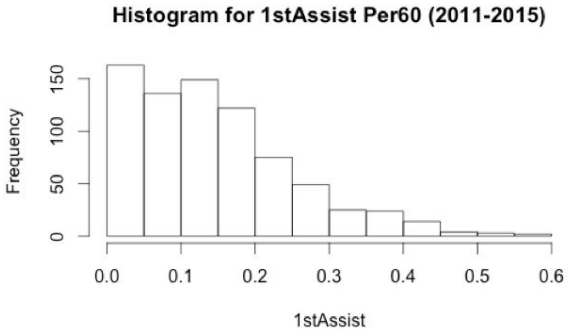


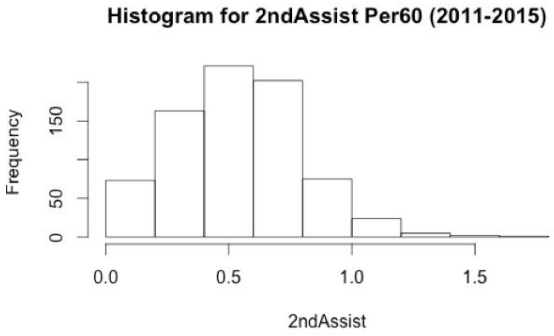


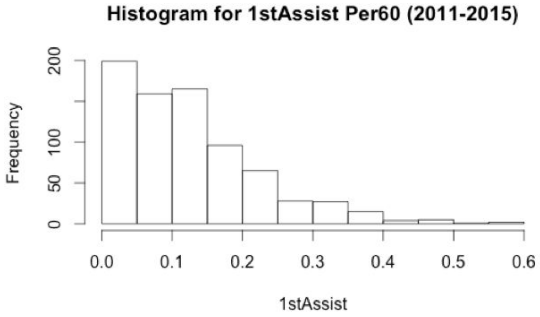
Appendix B: Summary of Goals and Assists

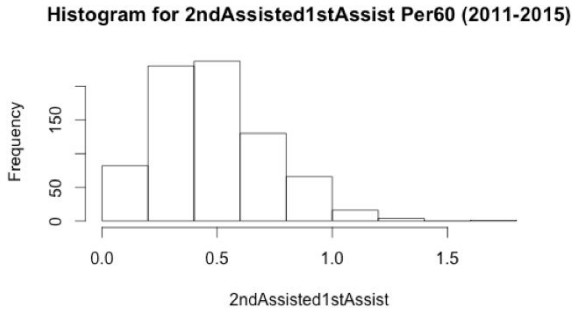
Independent Variable

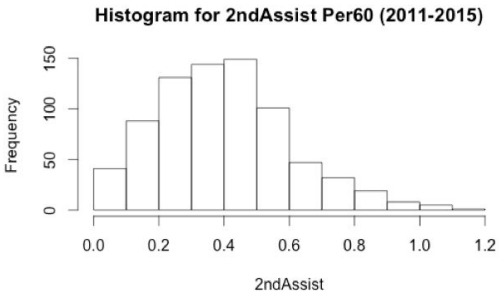


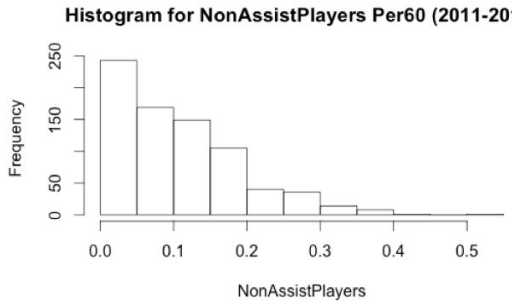


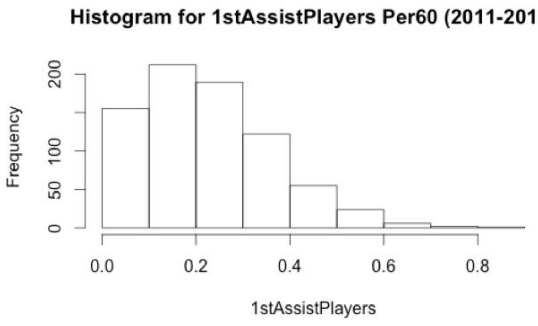


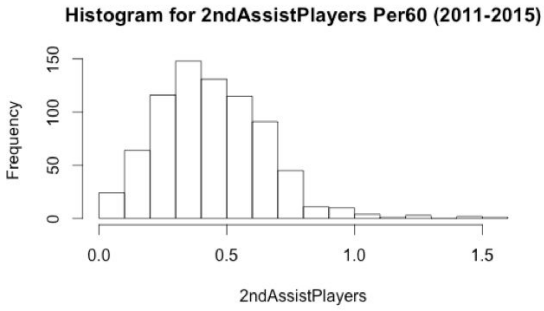












Dependent Variable