# NHL Corsi Statistic

The Effects of Simple Statistics on the Advanced Corsi Rating

This Capstone Project is submitted in partial fulfillment of the requirements for the course Data-Driven Decision-Making (MDA 620) during the Fall Semester of 2022.

While writing this Capstone Project, we have not witnessed any wrongdoing, nor have we personally violated any conditions of the LIU Honor Code.

Austin Rook

December 20, 2022

**Table of Contents**

Background (page 3)
Problem Scenario/Business Issue (page 5)
Objective/Goals for this Report (page 6)
Data Exploration (page 6)
Data Visualization (page 6)
Data Manipulation (page 8)
Methodology/Model Building/Analysis (page 8)
Model Selection (Page 11)
Conclusions/Recommendations (page 12)
Limitations (page 11)
Works Cited (page 14)

**Figures**

1. Highest and Lowest CF% Relationship Skater by NHL Team, 2019-2020 (page 4)
2. NHL Team Wins Against CF%, 2018-2019 (page 5)
3. Figure 3. Relationship Between Blocked Shots and Hits (page 7)

**Tables**

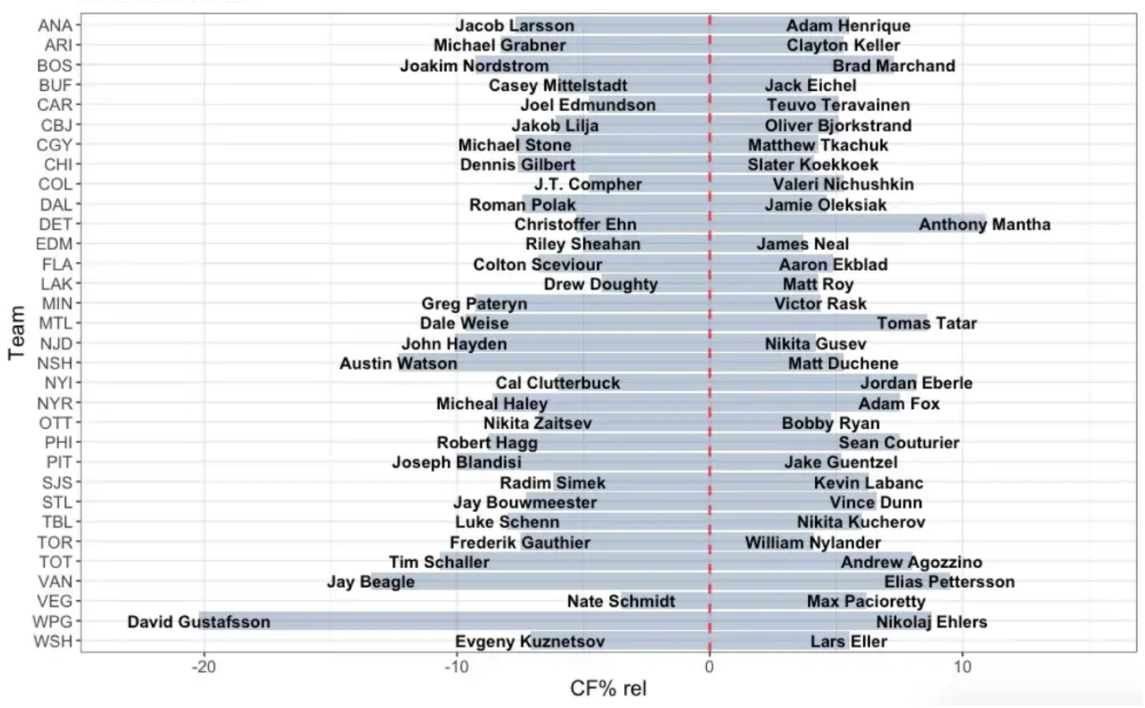
1. Summary Statistics (page 6)
2. oZS% Linear Regression Results (page 9)
3. FO% Linear Regression Results (page 9)
4. HIT Linear Regression Results (page 9)
5. BLK Linear Regression Results (page 9)
6. Multiple Regression Results (page 10)

**Background**

Many sports around the world have adopted advanced statistics to analyze individual and team performance. For example, Major League Baseball (MLB) has developed the "Wins Above Replacement" (WAR) which describes a player's value by analyzing how many more wins he would get compared to a replacement-level player. The MLB was the first league to adopt some of these advanced statistics but nowadays it is common practice across the major sports leagues. Over the past decade, the National Hockey League has developed four different advanced metrics for the sport. These statistics include Corsi, Fenwick, PDO, and Zone Starts. All of these metrics are taken when the two teams playing are at even strength, no powerplay or penalty kill statistics have an effect on these metrics. For the sake of this paper, I will only analyze one of these advanced statistics, the Corsi rating. "Corsi is determined by taking the number of shot attempts at even strength and dividing it by the number of shot attempts by the opponent. Shot attempts are different from shots on goal in that a shot attempt is any shot directed at the net. Shots on goal, shots that miss the goal, and blocked shots are added together to get the total shot attempts."[1] In Figure 1, one can see the highest Corsi-rated player compared to the lowest Corsi-rated player for every NHL team during the 2019-20 season. David Gustafsson of the Winnipeg Jets had the lowest Corsi rating while Anthony Mantha of the Detroit Red Wings had the highest.

[1] Masisak, Corey. "Hockey Advanced Statistics: What Is a Corsi Number?" *Sporting News*, 28 Oct. 2021.

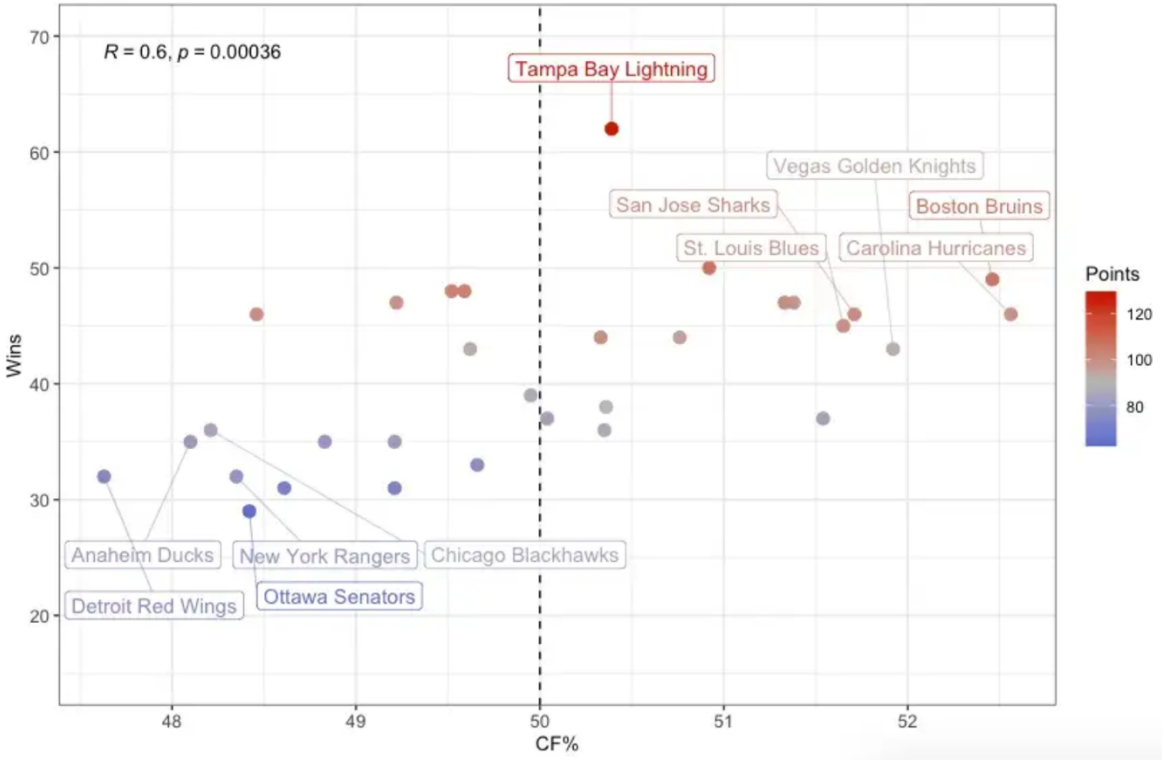
**Figure 1. Highest and Lowest CF% Relationship Skater by NHL Team, 2019-2020** [2]




As seen in Figure 1, the Corsi metric can be applied to individual players but also team performances. "A team's Corsi number takes the number of shot attempts by the team and divides it by the number of shot attempts by its opponent."[3] Corsi is the best approximation for puck possession, which analytic studies have shown is a great predictor of future success. In Figure 2, all 32 NHL teams are depicted along with their Corsi score, the number of wins they had, and the points they achieved in the 2018-19 season. The Tampa Bay Lighting achieved the highest Corsi rating, and the Ottawa Senators had the lowest.

[2] Lee, Christian. "Advanced Hockey Stats 101: Corsi (Part 1 of 4)." *Medium*, Hockey Stats, 6 Jan. 2021.
[3] Masisak, Corey. "Hockey Advanced Statistics: What Is a Corsi Number?" *Sporting News*, 28 Oct. 2021.

**Figure 2. NHL Team Wins Against CF%, 2018-2019** [4]




**Problem Scenario/Business Issue**

Just like many other sports, hockey is a dynamic sport. Many different aspects of the game can influence performance and the outcome of a contest. The Corsi rating system considers attempted shots but there are other factors that affect shot attempts. For example, if a team can win a faceoff in the offensive zone, they will have possession in the attacking zone. This may lead to more shot attempts due to higher levels of possession and increase the team's overall Corsi score. So, if a team is looking to increase their Corsi rating to gain success, what simple statistics should they analyze to increase this rating?

[4] Lee, Christian. "Advanced Hockey Stats 101: Corsi (Part 1 of 4)." *Medium*, Hockey Stats, 6 Jan. 2021.

**Objective/Goals of the Project**

The purpose of this paper is to analyze other simple statistics in the game of hockey to analyze their effects on a team's Corsi rating. These simple statistics include offensive zone starts, block shots, hits, and faceoff percentage. Through this analysis, I will be able to conclude whether a team should invest their efforts in increasing/decreasing these simple statistics to have a positive impact on their Corsi rating.

**Data Exploration**

To analyze the effects of other relevant statistics on the Corsi scoring system, I collected yearly data from the StatHead Hockey dashboard.[5]Data was collected from the 2007-08 season to the 2021-22 season. Statistics are a summary of season-long totals per team. Observations are representative of the best 100 teams from the specified data range in the category of Corsi For Percentage (CF%). From this data, one could observe the two best teams have the same Corsi For Percentage (CF%). With a score of 55.6, the Los Angeles Kings achieved this score in the 2013-14 season while the Florida Panthers achieved this score in 2021-22. Interestingly, the Los Angeles Kings appear four times in the top ten of Corsi ratings. This team won the Stanley Cup in 2011-12 and 2013-14.

**Data Visualization**

Table 1 provides the summary statistics (observations, mean, std. dev, min, and max) for each variable analyzed in this project. From Table 1, the standard deviation for CF%, oZS%, dZS%, and FO% is low because these are the top teams in the CF%, thus there is not much difference compared to teams at the bottom of these statistics.

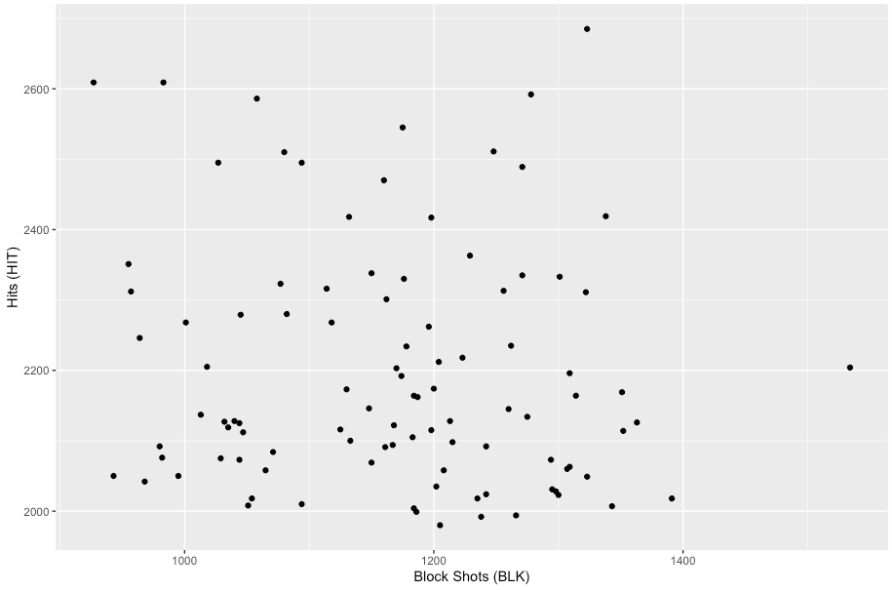
[5] "Team Advanced Stats Finder." *Stathead.com*.

**Table 1. Summary Statistics**

| Variable Names | Description | N | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|---|
| CF | Corsi For at Even Strength | 100 | 4624.05 | 305.16 | 3446 | 5374 |
| CA | Corsi Against at Even Strength | 100 | 4640.79 | 308.58 | 3939 | 5617 |
| CF% | Corsi For % at Even Strength (CF / (CF+ CA)) | 100 | 49.91 | 2.89 | 38 | 55.6 |
| oZS% | Offensive Zone Start % at Even Strength | 100 | 51.19 | 2.66 | 42.4 | 58.9 |
| dZS% | Defensive Zone Start % at Even Strength | 100 | 48.81 | 2.66 | 41.1 | 57.6 |
| FOW | Faceoff Wins at Even Strength | 100 | 2403.39 | 139.85 | 2061 | 2687 |
| FOL | Faceoff Loses at Even Strength | 100 | 2427.48 | 140.58 | 2107 | 2760 |
| FO% | Faceoff Win Percentage at Even Strength | 100 | 49.95 | 1.88 | 44.7 | 54.7 |
| HIT | Hits at Even Strength | 100 | 2199.19 | 170.76 | 1980 | 2685 |
| BLK | Blocks at Even Strength | 100 | 1167.72 | 121.41 | 927 | 1534 |

I thought it would be important to look at the number of blocked shots (BLK) and total hits (HIT) to see if there was any relationship between the two variables. From Figure 3, no relationship exists between these independent variables. Teams who play a more "physical" game tend to make more hits and sometimes that correlates an increase in willingness to block shots, but it seems there is no correlation between these two variables

Figure 3. Relationship Between Blocked Shots and Hits



**Data Manipulation**

During the data exploration stage of the process, I was lucky enough to come across data that had no missing value or irrelevant information. So, I didn't have to delete any variables or fill in gaps in the data. To obtain the relevant information for Table 1, I used RStudio to calculate the mean, standard deviation, min, and max of each variable used in this paper. RStudio also provided a platform to run the models included in this paper. Due to multi-collinearity issues, I had to drop Defensive Zone Start Percentage (dZS%).

**Methodology/Model Building/Analysis**

*Model 1*

For my first model, I decided to use a simple Linear Regression Model with each independent variable regressed on the dependent variable of Corsi For % (CF%). The equations for the following linear models are listed below. Beta ($\beta$) describes the coefficient for each independent variable in the model below. Tables 2-5 show the results of the Linear Regression Models.

$$CF\% = \beta_0 + \beta_1(oZS\%_{it})$$
$$CF\% = \beta_0 + \beta_1(FO\%_{it})$$
$$CF\% = \beta_0 + \beta_1(HIT\%_{it})$$
$$CF\% = \beta_0 + \beta_1(BLK\%_{it})$$

**Table 2. oZS% Linear Regression Results**

| | CF% |
|---|---|
| oZS% | 0.93*** |
| | (0.06) |
| Cons | 62.43 |
| Observations | 100 |
| R-Squared | 0.73 |

*Standard errors are in parentheses*
*** p<.01, ** p<.05, * p<.1*

**Table 4. HIT Linear Regression Results**

| | CF% |
|---|---|
| HIT | 0.002 |
| | (0.002) |
| Cons | 45.37 |
| Observations | 100 |
| R-Squared | 0.01 |

*Standard errors are in parentheses*
*** p<.01, ** p<.05, * p<.1*

**Table 3. FO% Linear Regression Results**

| | CF% |
|---|---|
| FO% | 0.63*** |
| | (0.14) |
| Cons | 18.48 |
| Observations | 100 |
| R-Squared | 0.17 |

*Standard errors are in parentheses*
*** p<.01, ** p<.05, * p<.1*

**Table 5. BLK Linear Regression Results**

| | CF% |
|---|---|
| BLK | -0.015 |
| | (0.002) |
| Cons | 67.03 |
| Observations | 100 |
| R-Squared | 0.38 |

*Standard errors are in parentheses*
*** p<.01, ** p<.05, * p<.1*

From Table 2, Offensive Zone Starts Percentage (oZS%) has a beta coefficient of 0.93 and is also significant. This coefficient means that a 1% increase in Offensive Zone Starts Percentage (oZS%) would result in a 0.93% increase in Corsi For Percentage (CF%). This linear model also has a relatively high R-Squared value of 0.73 compared to some of the other linear models. Similar to Offensive Zone Starts Percentage (oZS%), Faceoff Percentage (FO%) is significant and has a coefficient of 0.63 which means a 1% increase in Faceoff Percentage (FO%) would result in a 0.63% increase in Corsi For Percentage (CF%) (Table 3). The R-Squared of this specific model is 0.17 which is low compared to some of the other models. Tables 4 and 5 show the results for the linear models of Hits (HIT) and Block Shots (BLK) which aren't significant. For each model, an increase of one hit (HIT) would result in a 0.002%

increase in Corsi For % (CF%) and an increase of one blocked shot (BLK) would result in a negative increase of Corsi For % (CF%) by 0.015. These statistics might be low due to the number of shots and blocks in a given season. From Table 1, the range of hits (HIT) is 1980-2685 and blocked shots (BLK) is 927-1534. Thus, an increase of one hit or one blocked shot over the course of one season has a relatively small effect on other statistics.

*Model 2*

For the second model of this paper, I used a Multiple Regression Model. To examine the effects of relevant statistics on Corsi rating, I will use the CF% as my dependent variable for my regression model. My dependent variables will include Offensive Zone Start % at Even Strength (oZS%), Defensive Zone Start % at Even Strength (dZS%), Faceoff Win Percentage at Even Strength (FO%), Hits at Even Strength (HIT), Blocks at Even Strength (BLK). The equation of the model is listed below. Beta ($\beta$) describes the coefficients for the independent variables used in this paper. Table 6 shows the results from my multiple regression model.

$$CF\% = \beta_0 + \beta_1(oZS\%_{it}) + \beta_3(FO\%_{it}) + \beta_4(HIT_{it}) + \beta_5(BLK_{it}) + \varepsilon_{it}$$

**Table 6. Multiple Regression Results**

| | CF% |
|---|---|
| oZS% | 0.77*** |
| | (0.05) |
| FO% | 0.17* |
| | (0.07) |
| HIT | 0.001 |
| | (0.001) |
| BLK | -0.006*** |
| | (0.001) |
| Cons | 6.27 |
| Observations | 100 |
| R-Squared | 0.82 |

*Standard errors are in parentheses*
**** p<.01, ** p<.05, * p<.1*

From Table 6, one could observe that Offensive Zone Starts Percentage (oZS%) has the highest positive impact on a team's Corsi rating. I would interpret the coefficient of 0.77 as a one percent increase in a team's Offensive Zone Starts Percentage (oZS%) would result in a 0.77% increase in their Corsi For Percentage (CF%). This makes sense as an offensive zone start would increase the probability that a team will have higher shot attempts. Faceoff Percentage (FO%) has a coefficient of 0.17 which means it also has a positive effect on Corsi For Percentage (CF%). Unlike oZS%, FO% is not significant. If it was, I would conclude that a 1% increase in faceoff percentage would result in an increase of 0.17% increase in a team's Corsi For Percentage (CF%). Hits (HIT) was also not significant and its impact on Corsi For Percentage (CF%) is almost zero. The interpretation of this coefficient would be an increase of 1 hit would result in an increase of the Corsi rating of 0.001%. Finally, block shots (BLK) has a negative impact on Corsi For % (CF%). The coefficient is significant and would be interpreted as an increase of 1 shot would result in a 0.006 decrease in Corsi For % (CF%). As mentioned in Model 1, hits and blocked shots have very low coefficients because an increase of 1 hit/blocked shot over the course of the season should not have a large impact on other statistics. The R-squared value of this model is 0.82 which shows the model has a relative goodness of fit.

**Model Selection**

Analyzing the results from Model 1 and Model 2, I would select Model 2 as a better representation of factors that affect Corsi For Percentage (CF%). In a multiple linear regression model, each independent variable is considered when analyzing the dependent variable. Thus, our results are holistic and look at many factors that play a role in determining a team's Corsi For Percentage (CF%). In a simple linear regression model, only one independent variable is considered. While this may be valuable when looking at the effects of one variable, it lacks the

multi-dimensional approach of a Multiple Linear Regression model which is important in this specific analysis. The R-Squared value of Model 2 is 0.82 which is higher than any of the linear models in Model 1. This means that the independent variables included in Model 2 are explaining more of the independent variable compared to singular independent variables in Model 1.

**Conclusions/Recommendations**

Based on my results, I would recommend to NHL teams to increase their offensive zone starts and faceoff percentage while decreasing their blocked shots. If a team the run of play starts in the offensive zone, then the team is more likely to have a higher Corsi For rating which increases their Corsi For Percentage (CF%). An increase in faceoff percentage will result in more possession which will give a team greater opportunity to increase their attempted shots, thus increasing their Corsi For (CF). A greater faceoff percentage in the defensive zone will decrease a team's Corsi Against (CA). The combination of these two factors due to an increase in faceoff increase will result in a greater Corsi For Percentage (CF%). Summarizing the effects of blocked shots (BLK), if a team has more block shots then their Corsi For Percentage (CF%) will decrease.

**Limitations**

One large assumption this paper makes is higher levels of Corsi For Percentage (CF%) results in higher team success. A higher Corsi For Percentage (CF%) means that teams attempted shots is higher than attempted shots from the opponents. While shots may statistically mean there is higher chances of goals and thus wins, this is not always a great predictor of success. If a team has a goalie who is playing well, attempted shots may not result in a win. Another limitation of this paper is the simple statistics that were analyzed only gives a snapshot of a hockey game. The

independent variables used in this paper were the only ones available in the dataset. There are other statistics recorded in a hockey game such as 50/50 battles, passing accuracy, and offsides which may have an effect on the dependent variable analyzed in this paper.

**Works Cited**

Lee, Christian. “Advanced Hockey Stats 101: Corsi (Part 1 of 4).” *Medium*, Hockey Stats, 6 Jan. 2021, https://medium.com/hockey-stats/advanced-hockey-stats-101-corsi-part-1-of-4-29d0a9fb1f95.

Masisak, Corey. “Hockey Advanced Statistics: What Is a Corsi Number?” *Sporting News*, 28 Oct. 2021, https://www.sportingnews.com/us/nhl/news/what-is-a-corsi-number-explained-stats-nhl-hockey-advanced-statistics/9q6sdoe4l3o51jb3dm5lyjpxh.

“Team Advanced Stats Finder.” *Stathead.com*, https://stathead.com/hockey/tpbp_finder.cgi.