Austin Rook
Professor Singh
MDA 620
20 December 2022

NHL Corsi Statistic

Background

"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." (Masisak). See Figure 1 and Figure 2 for examples of individual and team Corsi ratings.

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

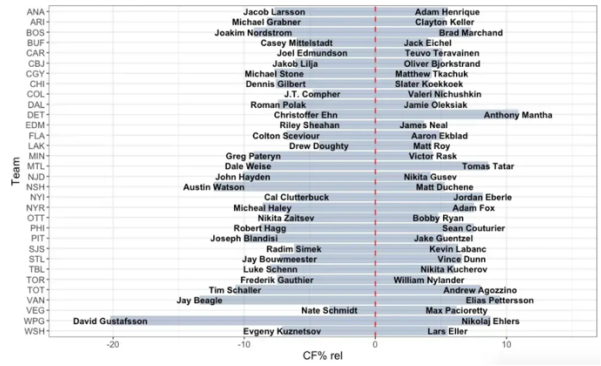
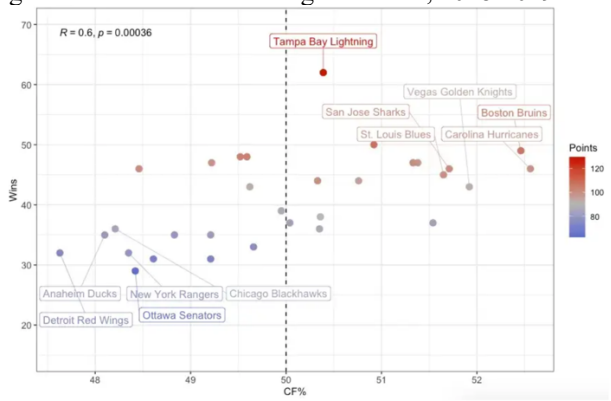


Figure 2. NHL Team Wins Against CF%, 2018-2019



Problem Scenario

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.

Objective

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

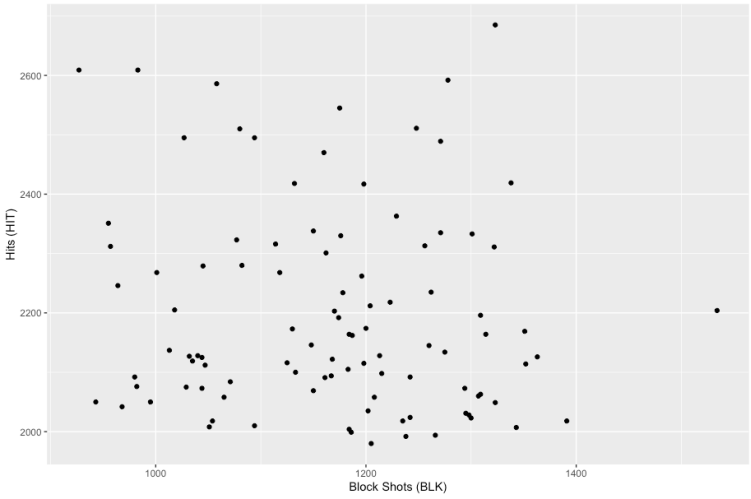
To analyze the effects of other relevant statistics on the Corsi scoring system, I collected yearly data from the StatHead Hockey dashboard. Data was collected from the 2007-08 season to the 2021-22 season. Statistics are a summary of season long totals per team. Observations are representative of the best 100 teams from the specified data range in the category of Corsi For Percentage (CF%).

Data Visualization

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

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 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 1 – Linear Regression Model

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 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 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 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

Model 2 – Multiple Regression Model

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*

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 the Multiple Linear Regression 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

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 a 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.