Visualizing NHL Statistics

An Analysis of 100 Years of Professional Ice Hockey Data

Ali Crump

December 13, 2019

Abstract

Sports analytics allow players and coaches to make more informed decisions by identifying where improvements in performance can be made. More and more teams at professional and even collegiate levels are incorporating statistical information into their practice and game plans, trying to get an edge on their competition in any way possible. The National Hockey League recently digitized and published more than 100 years of box score data, previously unavailable to the public. The website allows for the specification of any statistic for any year and team, for millions of data points. In this project, I create a web application using the recently published NHL data, aiming to identify patterns or tendencies of NHL teams and players. The application allows the user to visualize any statistic for their favorite team over the franchise's history or for a particular season. The application also creates a model to see if the Philadelphia Flyers 2019 - 2020 season is off to a good start. Finally, there is a page which allows for the visualization of any statistic for defunct NHL teams. Through the application we are able to come to a surprising conclusion that certain statistics like penalty kill percentage and poweplay percentage might not have as big of an impact on a team's success as perhaps once widely thought.

Background and Significance

Players and coaches constantly seek new ways of improving team and player performance. Sports analytics allows us to see the game in an objective and new way so that coaches and players can make accurate responses to enhance their game. In professional leagues, the impact of a team's performance goes beyond just winning a championship; oftentimes there are also financial and social implications to a team doing well, as better teams tend to draw more crowds and sell more tickets and merchandise. I find the impact of sports analytics extremely powerful, and for my project I looked at NHL data because I've always been fascinated by ice hockey. I grew up watching my brothers play and attending Flyers games with my family, so I was eager to analyze my favorite team. In my project I took years of NHL team and player data and created an online application which allows the user to visualize any team and any statistic that interests them. By visualizing the data using bar graphs, we can see trends and patterns that we might not have been able to see with just the raw data tables on the NHL's website.

Methods

In 2017, the NHL published its entire statistical history dating back 100 years to the first season in 1917. Data from 1987 to the present was previously available, but the NHL added more than 15 million data points from more than 24,000 game from the seasons between 1917 to 1987. The NHL's statistics team spent three years manually entering data from the original game sheets and another two years double-checking their work.

The data is not available for download so I had to web scrape it to use it for my project. Using a Google Chrome extension, I scraped the player data in eight different sections so that my computer could manage the massive amount of data. This process created eight different CSV files which I then bound together and analyzed in R. The scraping did not yield perfect CSV files, so I also had to manually set my variable types, remove duplicates, and remove NA and blank values.

Most of the variables in my project are standard ice hockey statistics that the NHL has tracked for the past 100 years like goals, assists, and points. More recent statistics include powerplay, penalty kill, and faceoff win percentages, for example, and those date back only a few decades. I kept all of these statistics as-is for my analyses.

I created a few variables in my Shiny App to be able to incorporate the user's inputs into my graphics. The user can input variables like a statistic, season, or their favorite team, and I renamed these inputs to make them easier to refer to in the creation of my plots. I also made the variable 'highlight,' which allows me to plot the Philadelphia Flyers' 2019-2020 season statistics on my model. 'Highlight' gives me a sense of how the Flyers are doing so far this season relative to teams in years past.

I first used a map to allow those unfamiliar with the NHL to get a sense of where NHL teams are actually located. The map shows the team and stadium name for each of their locations in North America. I then used bar charts to allow the user to visualize how a particular team or players from that particular team performed over the years or in a specific season. On the team statistics page, the user selects their favorite team and a statistic like points, shots, or powerplay percentage, for example, and they can see the progression of that team's performance on that statistic throughout the franchise's history. I used points as a measure of how well a team or player performed that season, so I colored the columns of my bar charts by the amount of points team or player finished the season with. The lightest shade of blue represents the season when the team had the best record, and the darkest shade the worst record. Similarly, the player page prompts the user to select a team, a statistic, and a season to visualize. The player page also includes bar charts colored by points and displays the fifteen players on that team who performed the best in the chosen statistic.

The models page displays a scatter plot and the linear regression of points versus the user's chosen statistic. Included on this page is the Flyers performance in the 2019-2020 season so far, shown as a vertical orange line. I only allowed the user to select statistics in the form of percentages and averages since the Flyers have only played a portion of their season.

The last page includes defunct NHL teams. I made this a new page because each team existed for different lengths of time and during different times in history, so it was difficult to include them in the team page. The defunct page is structured exactly like the current NHL team page, except some statistics are unavailable for older teams, occasionally resulting in blank bar charts.

Results

Not surprisingly, when we plot certain statistics like goals and wins, seasons in which the team had the most points also have the most goals and wins. We similarly see seasons in which the team had the most points having the least amount of goals against and losses. We see more interesting findings for statistics like powerplay percentage and penalty kill percentage. For seasons with the most points, we expect to see the highest percentages, but this is not often the case. For example, the Flyers' played their worst season in 2012-2013 with only 49 points, but they had their fourth highest penalty kill percentage that year, 85.9% (Appendix A). We can see similar surprising results for the Florida Panthers having their second highest powerplay percentage, 20.4% in their worst season (Appendix B).

The team page also allows us to see the general success of a franchise over the years. Looking at the Philadelphia Flyers again, we can see that the team had a long streak of success in the early 1970's leading up to their win of the Stanley cup (Appendix C). Similarly, this visualization tool allows us to see that the Pittsburgh Penguins faced a bit of a slump in the early 1980's (Appendix D). The player page produced similar findings; those with the most points most expectedly had the most goals and assists. Players with the highest number of penalty minutes were not typically the players with the most points. A more surprising finding is that players with less points often had the highest shooting percentage (Appendices E, F). For my models, I found that statistics like shots per game played, faceoff win percentage, powerplay percentage, our prediction for the amount of points the Flyers might finish the 2019-2020 season with is very high. For the other statistics like goals for, goals against, and shots against per game played, the prediction for points was much lower (Appendices G, H). For the defunct teams page we have similar relationships between points and the different chosen statistics, but the results are more difficult to visualize and interpret. Teams like the Brooklyn Americans or the Montréal Wanderers only had one season, and the documentation of certain statistics like penalty kill percentage are more recent, so the data does not allow for much interpretation.

Conclusion

My goal for my project was to use my visual analyses to see patterns or tendencies that were not originally obvious. On the team page, we saw that the highest powerplay and penalty kill percentages did not always correspond to the seasons with the most points. This finding suggests that teams should spend more time practicing with even-strength than practicing man-up man-down situations. This conclusion was not as clear when we just viewed the raw data on its own.

Through the visualizations, we were also able to see more clearly when a team saw a lot of success and when a team was going through a bit of a slump. This visualization could allow team owners and managers to catch onto slumps and nip them in the bud by proactively making adjustments for the next season.

Our analysis on the player page allows us to see that some players with the least amount of points had the highest shooting percentages. The players who had the highest shooting percentages might have also taken the least amount of shots, which could explain this finding. With further investigation, though, this result might suggest to coaches that certain players who might not be considered the strongest attackmen should be shooting more.

By plotting the Flyers' 2019-2020 statistics on the regression models, we see that the Flyers performed outstandingly well in terms of their shots per game played and faceoff win and powerplay percentages. While statistics like goals per game might better inform us about how the Flyers will perform this season, it was still interesting to see that the Flyers are performing particularly well in these statistics not just compared to other teams this season but to most other teams in the history of the NHL.

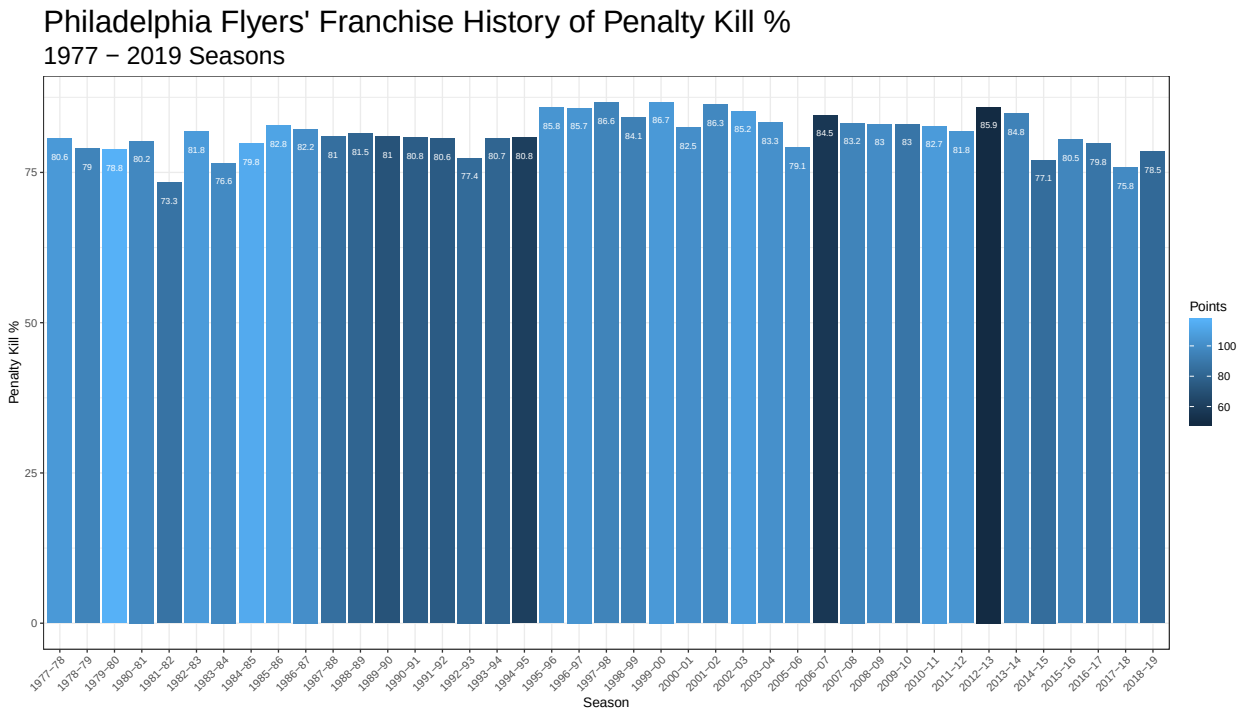
A signficant limitation of the project is that you cannot compare the statistics of teams to each other for a particular season and you cannot compare the statistics of players from different teams. In the future I would like to try to add a feature that allows the user to compare statistics across multiple NHL teams.

Another limitation is how I measured the success of a team or player. I defined success as the number of points a team or player finised the season with, but the team that wins the Stanley Cup is not always the team with the most points, and the best player is not necessarily the one who scores the most goals or has the most assists. In the future I might include some sort of final playoff ranking to measure success and I might try to create different features for different postition. Finally, in the future I'd like to include statistics for goalies, which do not appear in this project.

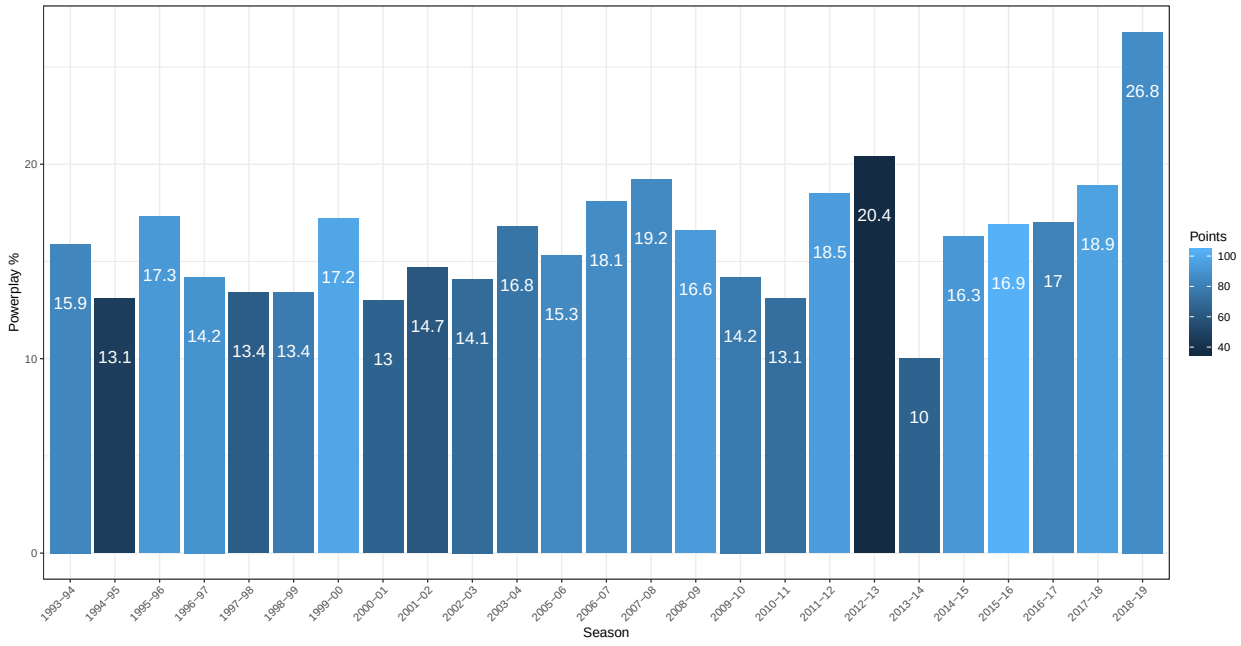
By analyzing statistics like powerplay percentage and penalty-kill percentage over time, coaches might adapt practice plans to include more even-strength hockey. Coaches might instruct players with high shot percentages to shoot more. Ultimately, this application allows us to investigate NHL teams' and players' statistics further through visualizaiton to inform coaches' and players' decision-making. Through the application we can see trends in any franchise's history for any statistic recorded dating back more than 100 years.

Appendix

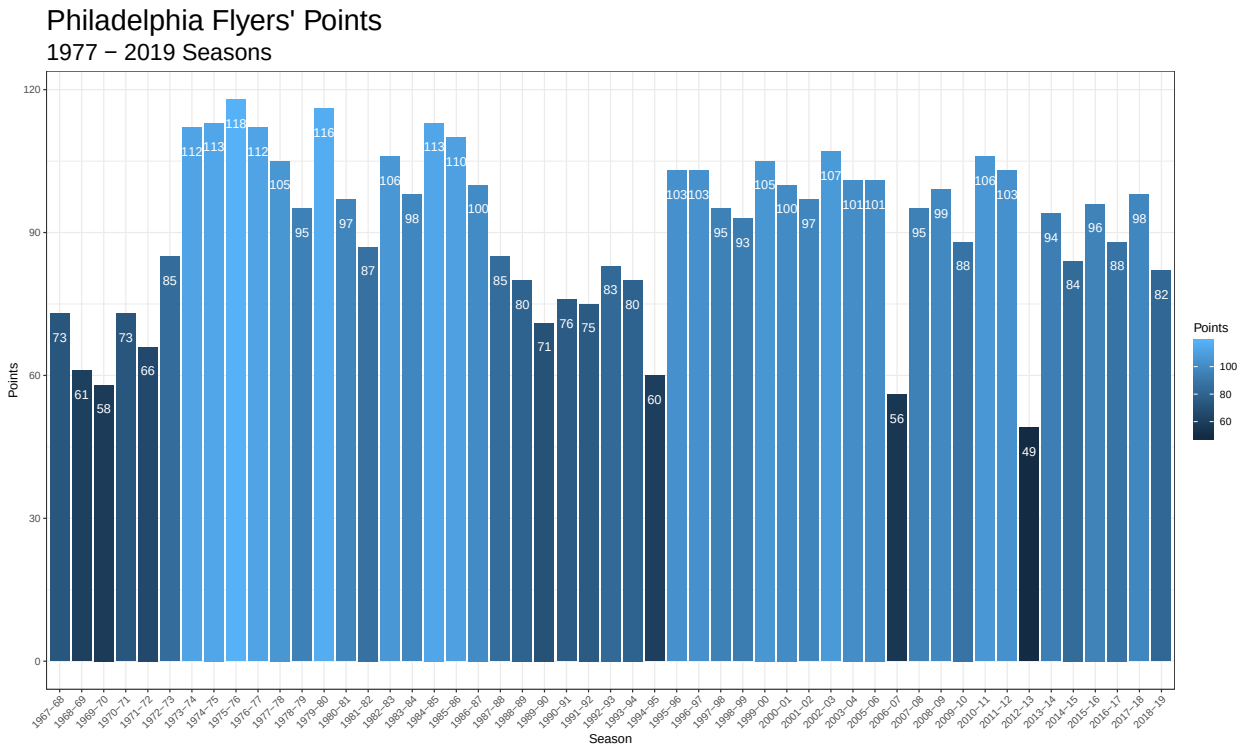
A


Philadelphia Flyers' Franchise History of Penalty Kill %
1977 − 2019 Seasons
Penalty Kill %
Season
Points
100
80
60
0
25
50
75
80.6
79
78.8
80.2
73.3
81.8
76.6
79.8
82.8
82.2
81
81.5
81
80.8
80.6
77.4
80.7
80.8
85.8
85.7
86.6
84.1
86.7
82.5
86.3
85.2
83.3
79.1
84.5
83.2
83
83
82.7
81.8
85.9
84.8
77.1
80.5
79.8
75.8
78.5
1977−78
1978−79
1979−80
1980−81
1981−82
1982−83
1983−84
1984−85
1985−86
1986−87
1987−88
1988−89
1989−90
1990−91
1991−92
1992−93
1993−94
1994−95
1995−96
1996−97
1997−98
1998−99
1999−00
2000−01
2001−02
2002−03
2003−04
2005−06
2006−07
2007−08
2008−09
2009−10
2010−11
2011−12
2012−13
2013−14
2014−15
2015−16
2016−17
2017−18
2018−19

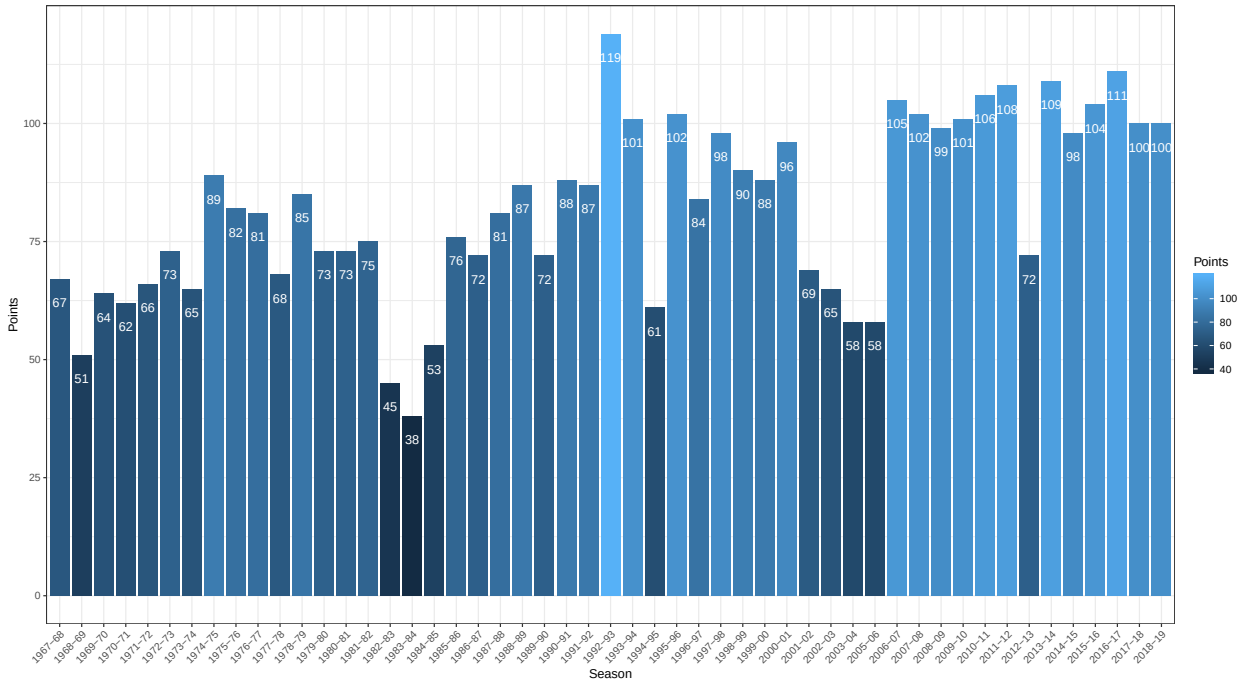
B


Florida Panthers' Franchise History of Powerplay %
1993 − 2019 Seasons
Powerplay %
Season
Points
100
80
60
40
0
10
20
15.9
13.1
17.3
14.2
13.4
13.4
17.2
13
14.7
14.1
16.8
15.3
18.1
19.2
16.6
14.2
13.1
18.5
20.4
10
16.3
16.9
17
18.9
26.8
1993−94
1994−95
1995−96
1996−97
1997−98
1998−99
1999−00
2000−01
2001−02
2002−03
2003−04
2005−06
2006−07
2007−08
2008−09
2009−10
2010−11
2011−12
2012−13
2013−14
2014−15
2015−16
2016−17
2017−18
2018−19

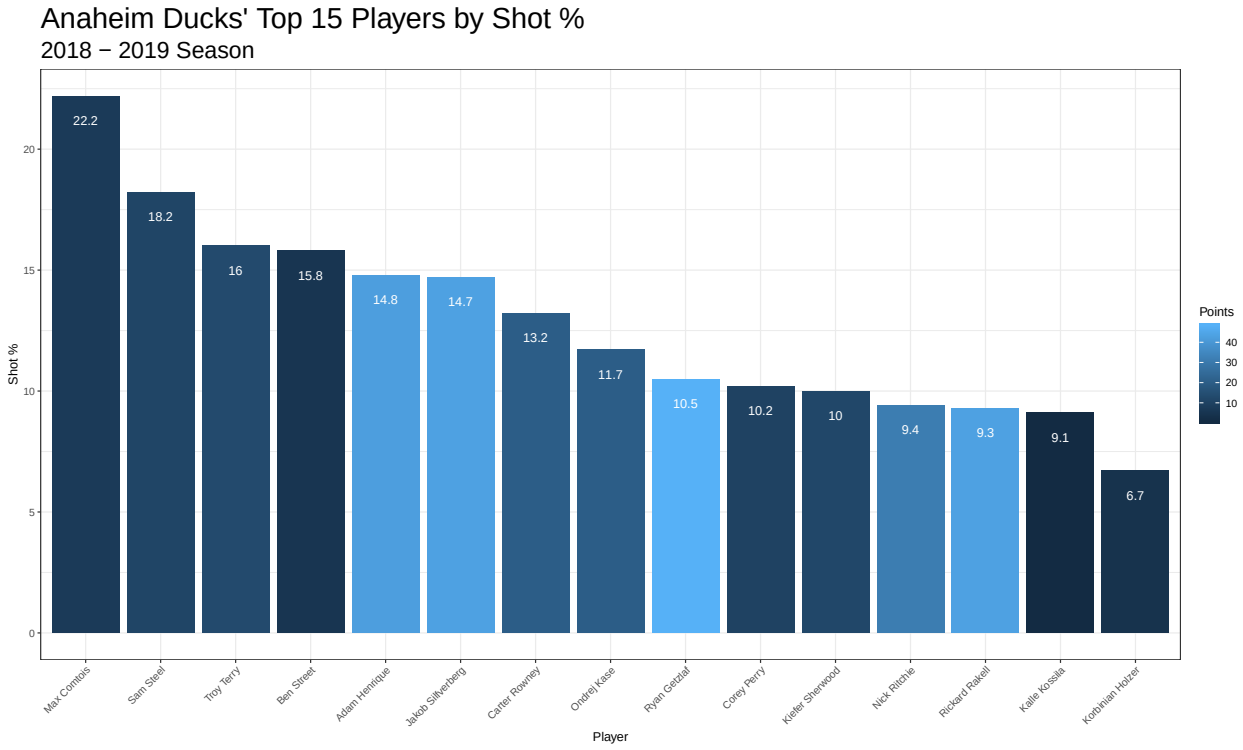
C


Philadelphia Flyers' Points
1977 − 2019 Seasons
Points
Season
Points
100
80
60

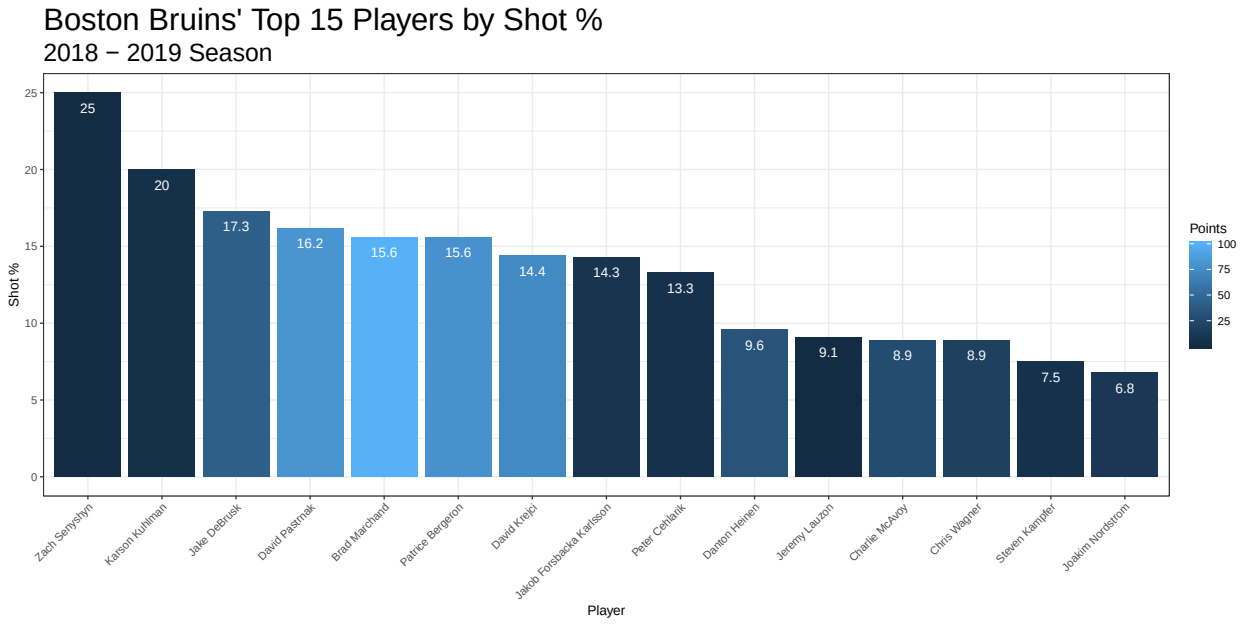
D


Pittsburgh Penguins' Points
1967 − 2019 Seasons
Points
Season
Points
100
80
60
40

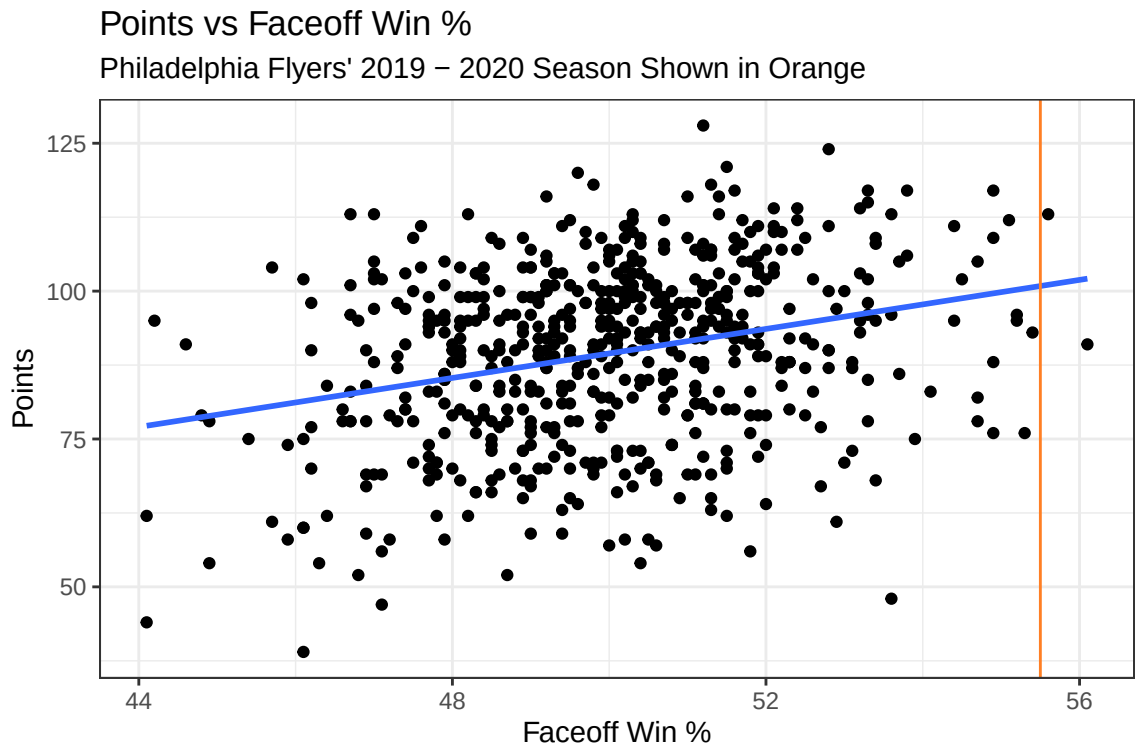
E


Anaheim Ducks' Top 15 Players by Shot %
2018 − 2019 Season
Shot %
Player
Points
22.2
18.2
16
15.8
14.8
14.7
13.2
11.7
10.5
10.2
10
9.4
9.3
9.1
6.7
Max Comtois
Sam Steel
Troy Terry
Ben Street
Adam Henrique
Jakob Silfverberg
Carter Rowney
Ondrej Kase
Ryan Getzlaf
Corey Perry
Kiefer Sherwood
Nick Ritchie
Rickard Rakell
Kalle Kossila
Korbinian Holzer

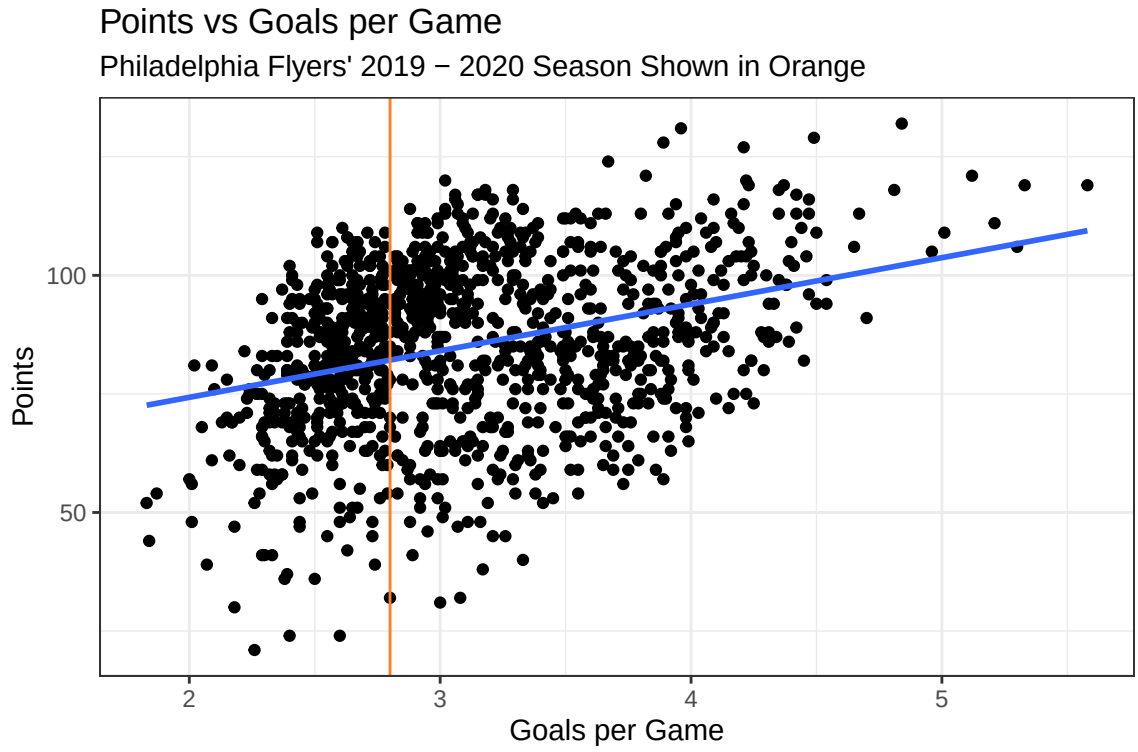
F


Boston Bruins' Top 15 Players by Shot %
2018 − 2019 Season
Shot %
Player
Points
25
20
17.3
16.2
15.6
15.6
14.4
14.3
13.3
9.6
9.1
8.9
8.9
7.5
6.8
Zach Senyshyn
Karson Kuhlman
Jake DeBrusk
David Pastrnak
Brad Marchand
Patrice Bergeron
David Krejci
Jakob Forsbacka Karlsson
Peter Cehlarik
Danton Heinen
Jeremy Lauzon
Charlie McAvoy
Chris Wagner
Steven Kampfer
Joakim Nordstrom

G


Points vs Faceoff Win %
Philadelphia Flyers' 2019 − 2020 Season Shown in Orange
Points
125
100
75
50
44
48
52
56
Faceoff Win %

H


Points vs Goals per Game
Philadelphia Flyers' 2019 − 2020 Season Shown in Orange
Points
100
50
2
3
4
5
Goals per Game