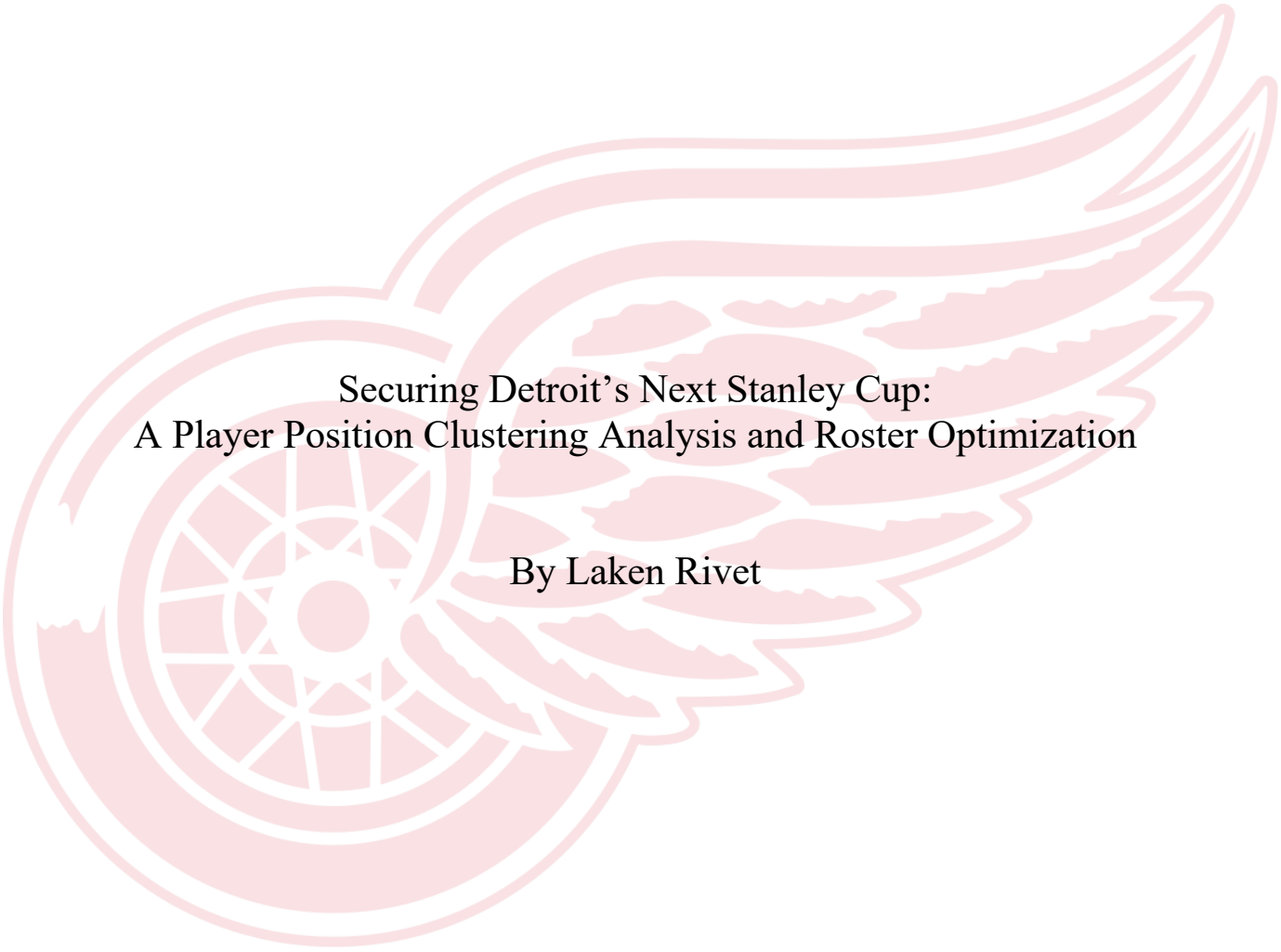


Securing Detroit's Next Stanley Cup:
A Player Position Clustering Analysis and Roster Optimization

By Laken Rivet

Table of Contents

Introduction

Ice hockey, like many other professional sports, contains several positions in which a player can be classified. These positions are derived from where the player is located on the ice and whether they're on offense or defense. The most that can be learned from a player's position is their handedness for shooting, and in the case of forwards, that they are the most likely to take a faceoff. There are certain assumptions about the offensive and defensive positions like the notion that forwards typically score more while defensemen hit more often. But what if we encounter players who break these stereotypes? Will their position falsely imply their strengths and weaknesses? To truly understand a player's style of play, new classifications need to be created that identify the unique qualities of a player. How to create these classifications and turn them into actionable insights is the problem this project seeks to solve.

The primary goal of this study is to provide the Detroit Red Wings with player classifications that are representative of play style which can be used for scouting, acquisition, or internal evaluation. To do so, a machine learning method called clustering was performed on basic and advanced player statistics for forwards and defensemen. The clustering algorithm then identified natural groups, or clusters, of players based on similarity. With this information, examination of the current team was performed as well as a roster optimization to identify the ideal roster heading into next season. The results of clustering and roster optimization were used to compile a set of recommendations to guide the Red Wings in potential roster changes. It is the hope that the classifications and recommendations provided will contribute to elevating the team to playoff qualification in the upcoming season.

Throughout the remainder of this report, additional details of the analyses will be provided. First, a literature review will be performed on similar, previous studies and how they informed the approach to the current study. The methodology behind the processes of data acquisition, data cleaning and filtering, principal component analysis, clustering, and constrained optimization will then be outlined. Next, the results of the clustering analysis will be presented in addition to an evaluation of the composition Detroit's current roster. Two optimal rosters and the moves taken to create them as determined by constrained optimization will follow. In the conclusion of the report, key findings will be summarized, and recommendations will be presented. Finally, the limitations of the analyses, future research to address said limitations, and next steps for initiating future research are discussed. With expectations for this analysis established, previous work to address the player classification problem can be considered.

Literature Review

The use of clustering to identify player classifications is well documented for and largely began with the National Basketball Association (NBA). Players are classified into 5 positions based on where they typically preside on the court. Basketball positions also lack the ability to encapsulate playing style. As such, there is a plethora of studies that have been performed with the goal of classifying players into groups that reflect their strengths and weaknesses, many of which utilize k-means clustering to do so (Hussain 2019; Stern n.d.; Batra 2022). The reasoning provided for the use of the k-means clustering algorithm in particular is its simplicity in implementation and effectiveness in grouping similar players. Given the many similarities between the NBA and NHL – the number of positions, high-speed play, frequency of passing, and foul/penalty occurrence – using the same approach to cluster hockey players should translate seamlessly.

A study by Lähevrita (2018) examining the effectiveness of clustering in classifying NHL players by position found that claim to be true. It should be noted that the study only considered forwards and defensemen, as goalies are not comparable to skaters. Lähevrita input six statistics standardized per 60 minutes of on-ice time including goals, assists, shots, hits, blocks, and corsi rating, as well as average time on ice per games played into the k-means clustering algorithm. The number of clusters to be output was set to 2, 3, 4, and 5 and each set of resulting clusters were evaluated by player position makeup. Lähevrita found that with the addition of clusters, the algorithm was able to separate forwards and defensemen increasingly well. While Lähevrita's (2018) results were informative in proving that clustering can differentiate subcategories of forwards and defensemen, further research was performed to find studies that performed additional analysis on the clusters themselves to identify unique playing styles.

Two such studies were identified and reviewed at length. The first, by Schulte et al. (2017), utilized proprietary on-ice location data to create player clusters. Their data recorded every event in every game of the 2015-2016 season as well as the player responsible for the event and their location on the ice during said event. Once players were clustered by locational data, they were ranked within their clusters by a stat engineered by the authors called scoring impact. To cluster the players, they used the affinity propagation algorithm and found it created 4 clusters for defensemen, 4 clusters for forwards, and 1 for goalies. As with Lähevrita (2018), they found clustering did well in separating clusters by position.

The second study, authored by Stimson (2017), utilized proprietary passing data to create player clusters. His data considered approximately 900 games of the 2015-2016 season and only considered the offensive passing style of players. Prior to feeding the data to the clustering algorithm, Stimson performed feature engineering to create standardized indexes which measured different aspects of passing like controlled zone entry assists and individual shots. Like Lähevrita (2018), his features were standardized per 60 minutes of on-ice time. Stimson also separated forwards and defensemen into two data sets prior to clustering. He used a k-means clustering algorithm and found the ideal number of forward and defensemen clusters to be 4. Stimson concluded his study by considering the optimal combination of player types on defensive and offensive lines.

The methods and results of relevant studies were used to inform certain decisions made in this analysis. Despite both Schulte et al. (2017) and Stimson (2017) using proprietary data, the decision was made to use only publicly available data for this project. This was partially due to lack of access to such data but also because a goal of this project was to create clusters on data that is accessible by anyone. Inspired by Stimson's (2017) success in separating forwards and defensemen prior to cluster analysis, the same was done for this study. Both Lähevrita (2018) and Stimson (2018) standardized variables by 60 minutes of on-ice time. A similar method was performed in this analysis in that player statistics, excluding percentages, were divided by the player's total time on ice. Again, this was inspired by success found but previous study but also with two considerations in mind. First, as the 2020-2021 season was reduced by COVID-19, total stats from that year couldn't be fairly compared to other seasons, requiring a form of standardization. Second, dividing stats by total time of ice gives a clearer picture as to how a player utilizes their time on ice regardless of how much they are given. After all relevant decisions were made regarding the data to be used in the study, data collection began.

Methods

Data Acquisition

Player performance data was acquired from hockey-reference.com. Two separate sets of player statistics – basic and advanced – were collected for each player for the 2020-2021, 2021-2022, and 2022-2023 NHL seasons. Three seasons worth of data was collected to ensure that upon data cleaning and filtering, there would be a large enough sample size to extract meaningful results. Data was not presented in a downloadable format and thus required web crawling and scraping to extract from the internet. To do so, the "read_html", "html_nodes", and "html_text" commands from the rvest R library were utilized. First, a Cascading Style Sheet (CSS) Selector called SelectorGadget was used to read the website's HTML page and identify the tags associated with the text to be acquired. Then a for loop was written to iterate through the relevant website identified in the "read_html" command. The for loop used the "html_nodes" command to locate the necessary data with the tags from SelectorGadget. Finally, the "html_text" command was used to transform the data into text that could be saved into a data frame in R. This process was repeated for both the basic and advanced statistics pages, and they were saved into two data sets, respectively.

Active player and free agent data were obtained from capfriendly.com. Active player data was collected for the 2022-2023 season only and free agent data was collected for players entering free agency in the 2023-2024 season. Once again, the data was not in an easily extractable format, so web scraping and crawling was performed on 32 web pages using the same tools and techniques listed above. Only basic stats and contractual information – player name, age, team, games played, goals, assists, points, points per games played, plusminus, shots, shooting percentage, average time on ice, handedness, salary, contract clause, agency, and cap hit – were collected for active players. For the free agent data, only the names of the free agents were collected and stored in a list format as they would be used for data filtering.

Data Cleaning and Filtering

Once both player performance data sets were acquired, the same cleaning operations were performed on each. First, variable names were cleaned utilizing the "clean_names" function from the janitor library. Certain variables were renamed after cleaning for clarity. Next, the data sets were filtered for entries that did not contain a value and replaced with "NA" values. Variables containing information regarding time on ice in the minute to second format (i.e., 09:30) were split into minutes and seconds variables and then recombined into a singular time on ice in minutes variable by adding minutes to seconds divided by 60. All columns containing numerical variables were then converted into a numeric class. The next step of the data cleaning process was to create an additional player position variable that contained only "F" or "D", forward and defensemen, respectively. Next, the standardized variables were created. This included dividing all numeric raw statistics, not including percentages, by total minutes on ice for each respective player.

It should be noted that players who were traded within the season have two entries in the 2020-2021 season data sets, one for each team. Players who were traded in the 2021-2022 and 2022-2023 seasons have three entries, one for each team and a total for the entire season between both teams. The decision to include all entries was made due to the fact that playing styles could change from one team to another and may contribute to the analysis.

The two data sets were then merged on the player, team, position, season, games played, and age variables. The merged data set contained both basic and advanced skater stats. Finally, the merged data set was split into two data sets, forwards and defensemen, by the additional player position variable.

As defensemen rarely attempt faceoffs, variables relevant to faceoffs in the data set were examined. It was found that the maximum number of faceoffs any defensemen took in the entire data set was 1. Thus, three variables – faceoff wins, faceoff losses, and faceoff percentage – were removed from the defensemen data set.

Prior to cluster analysis, both data sets were filtered further to reduce outliers that could potentially skew results. The primary variable considered for filtering was games played as few games played can heavily impact statistics like points per game. To determine a cutoff point for inclusion in the analysis, a histogram and a density plot of games played were created for each data set and can be observed in Figures 1 and 2.

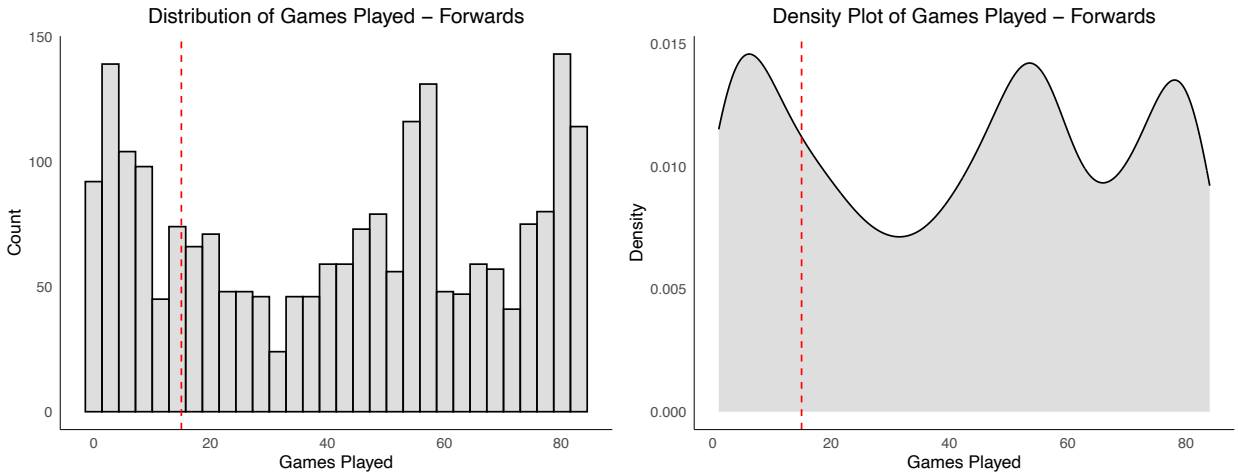


Figure 1. Histogram of games played (left) and density plot of games played (right) for forwards in the 2020-2021, 2021-2022, and 2022-2023 NHL seasons.

The first quartile of the data, or the cutoff for the bottom 25% of games played, was calculated to be 15 for forwards. The red dashed line in both plots of Figure 1 illustrates where the division occurred. As cutting off at this point retained approximately 75% of the original data set, the forwards data was filtered to only include players who played more than 15 games in a season. This led to the reduction of entries from 2,184 to 1,632.

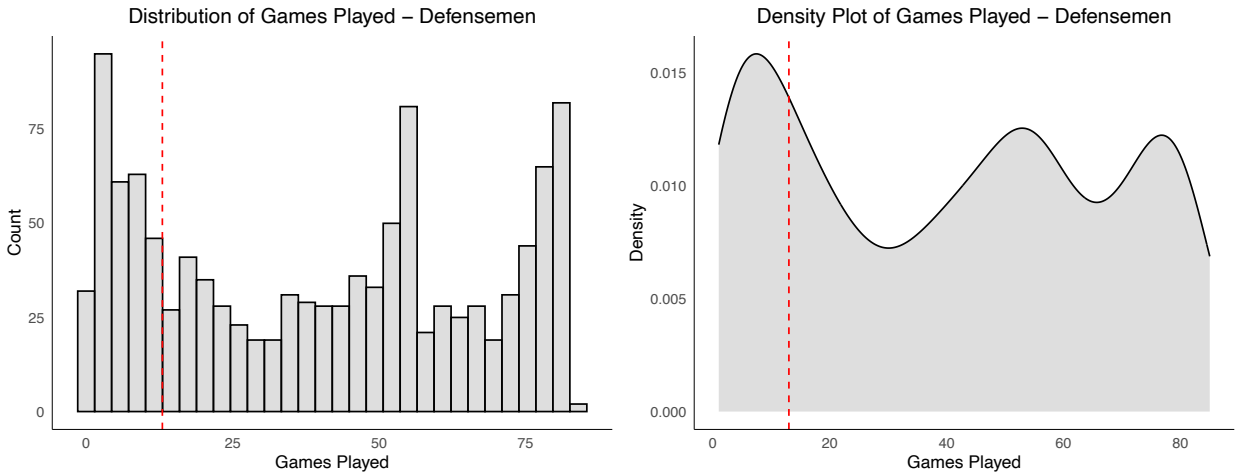


Figure 2. Histogram of games played (left) and density plot of games played (right) for defensemen in the 2020-2021, 2021-2022, and 2022-2023 NHL seasons.

The first quartile of the defensemen data set was calculated to be 13. Again, the red dashed line in both plots of Figure 2 illustrates where the division occurred. Due to the fact that the defensemen data set had less entries to begin with, the data was filtered to only include players who played 13 or more games in a season. Including those with exactly 13 games in addition to those with more allowed for a retention of approximately 76% of the original data set. In total, the defensemen data set was reduced from 1,150 entries to 870 entries.

The final step in cleaning and filtering the player performance data was checking both data sets for "NA" values. Upon completing this check, an issue with the 2020-2021 data was revealed. It appeared that for all players who played on multiple teams that season, the variables that represent total shots attempted and percentage of shots that made it on net were not properly summed in the player's total row for all teams. To remedy this, for loops were written for each data set that looked up the player by name and season, then proceeded to sum total shots and average percentage of shots that made it on net across all teams the player was on. Once this was completed the data sets were checked for "NA" values again and revealed 3 in total, 1 forward and 2 defensemen. Further exploration of these players showed that they each only had a total row but not individual team rows, likely an error on the website's part during the accumulation of data. Due to this error, the missing columns could not be imputed as there was no additional statistics from that season to draw from. As such, all 3 players were dropped from their respective data sets. A data dictionary that provides the variables included in the player performance data sets can be found in Appendix A.

The active player and free agent data required very little cleaning. In the active player data, the average time on ice variable was converted from the minutes to seconds format to minutes only using the same procedure performed on the player performance data sets. Financial variables like salary and cap hit were cleaned to remove dollar signs and commas, then converted to the numeric class. As the free agent data was stored in a list format, each element of the list was iterated through to remove all additional characters besides the player's name. Upon completion of this step, the active player data was then filtered to include on those players whose names are the in the free agent list and Detroit Red Wings players. This data set would later be merged with the clustering results and then fed to the constrained optimization solver. With that,

all data was prepped for analysis, and work on the principal component analysis (PCA) could begin.

Principal Component Analysis

Prior to PCA, correlation matrices were created for both data sets and can be found in Appendix B. As many variables exhibited strong, linear relationships, there was justification for using principal component analysis for two reasons. First, highly correlated variables will skew clustering results, whereas performing PCA outputs principal components that are uncorrelated to one another and thus will not impact clustering (Steiger 2015). Second, by reducing the number of inputs into the clustering algorithm, the computing power and time required to perform the task will be reduced. For these reasons, PCA was used in this project.

The first step in performing PCA was scaling the numeric variables to be used in clustering so that every variable had a mean of zero. This was done using the built-in "scale" function in R. Then PCA was performed on the scaled data using the "prcomp" function from the stats library. This resulted in the computation of 40 and 37 principal components for the forward and defensemen data set, respectively. In accordance with the Kaiser-Guttman rule, principal components with eigenvalues greater than 1 were kept for clustering. The rule states that a principal component with an eigenvalue larger than 1 represents an "above average" component, and thus should be kept for analysis (Steiger 2015). Scree plots, which display the eigenvalues of the first 15 components were created for both data sets and can be found in Figure 3.

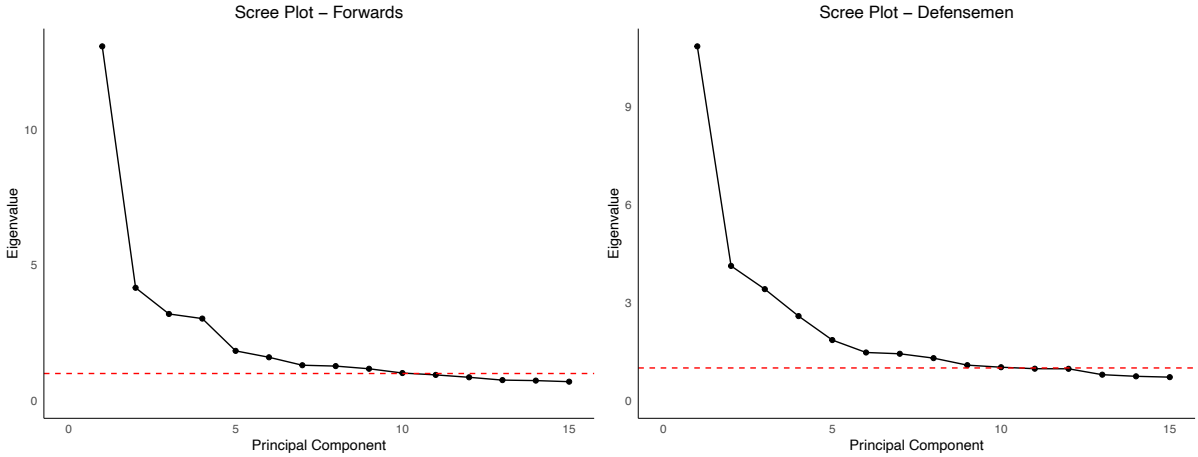


Figure 3. Scree plots of the first 15 principal components for forwards (left) and defensemen (right) in the 2020-2021, 2021-2022, and 2022-2023 NHL seasons.

For forwards, the first 10 principal components had an eigenvalue higher than 1. Together the components that were kept for clustering explained 79.2% of variance within the scaled data set. For defensemen, the first 10 principal components were also found to have an eigenvalue higher than 1. The defensemen components that were kept for clustering explained slightly less variance than the forward components at 78.8%.

Clustering

With the principal components created and the ideal number to keep in the analysis determined, the data was ready to be input into the clustering algorithm selected for this project, k-means. The decision was made to use k-means because it's the most effective at creating

similarity-based clusters while also considering the data set as a whole compared to other clustering algorithms such as hierarchical or density-based (ODSC 2018). Additionally, the data for this project is most similar to the data used by Stimson (2017), so it was logical to utilize the same algorithm he found success with.

K-means is a very powerful algorithm but has one major drawback – the number of clusters to partition the data into must be determined by the user. Choosing the optimal number of clusters is a challenge as there are many tests that can be performed to aid in the decision, but they often provide differing results. To approach this issue, similar studies were first considered. Shulte et al. (2017) used the affinity propagation clustering algorithm, which automatically selects the numbers of clusters to partition the data into. They found the algorithm created 4 clusters for forwards and 4 clusters for defensemen. Stimson (2017) also found the ideal number of clusters for both forwards and defensemen to be 4.

With these similarities between the previous in mind, five types of tests were run to estimate the optimal number of clusters for the data. The first is widely known as the elbow method, in which a plot of the total within sums of squares against the number of potential clusters is used to visually identify the number of clusters where the values level out. The second calculates the average silhouette, or average distance between clusters to identify the number of clusters at which the average silhouette score is the highest. The third considers intra-cluster variation compared to expected values, also known as the gap statistic, to identify the number of clusters that maximizes the gap statistic. The fourth test utilized hierarchical clustering, as it can determine the ideal number of clusters without user inputs. The final test used the "NbClust" function from the NbClust R package. The function calculates 30 indices, or methods to determine cluster number, and reports the number of clusters that the majority of methods suggested. The resulting cluster recommendations from all tests can be observed in Table 1.

Table 1. Resulting number of clusters for forwards and defensemen.

Method	Number of Forward Clusters	Number of Defensemen Clusters
Elbow Method	4	3
Average Silhouette	2	2
Gap Statistic	4	3
Hierarchical Clustering	3	3
NbClust	2	2

For the forwards data set, the cluster number recommendations were tied at 2 and 4. As results from the relevant studies found 4 clusters to be ideal for forwards, 4 clusters were used for this project. The defensemen data set provided clearer results with the majority of the methods determining 3 clusters to be optimal for defensemen. Thus, defensemen were divided into 3 clusters.

K-means clustering was performed on both data sets using the "kmeans" function from the stats R library. The default arguments of the function were used besides specifying the number of clusters for each data set, 4 and 3, respectively, as well as the number of starts, which was set to 25. After clustering, the final machine learning method to be used was constrained optimization.

Constrained Optimization

Constrained optimization is the process of maximizing or minimizing an objective function under a set of conditions, or constraints. As roster construction is binary in nature – either a player is added to the roster or not – it falls under a category of optimization called a knapsack problem. In such problems, the goal is to identify the combination of items, or in this case players, that create the most value in a limited number of spots. Linear integer programming is well-suited to solve knapsack problems as it can consider all possible combinations of players and their accompanying values. It is widely considered the best method to approach the knapsack problem, and thus was used for this project (Leite 2022).

To create the optimal roster, the "lp" function of the lpSolve R library was utilized. The objective function was defined as total points per games played for the team with constraints regarding the number of Red Wings players, the number of forwards, the number of defensemen, and all resulting player cluster types. The default arguments of the function were used besides specifying that all variables were binary (all.bin = TRUE). Several combinations of player cluster constraints were attempted, and their resulting success is compared in the Results section. First however, the clusters themselves must be examined.

Results

Forward Clusters

To understand how forwards were clustered and the defining characteristics of each player cluster, it's best to see their general patterns. Figure 4 provides information about where each cluster's center, or mean, falls in comparison to the other clusters.

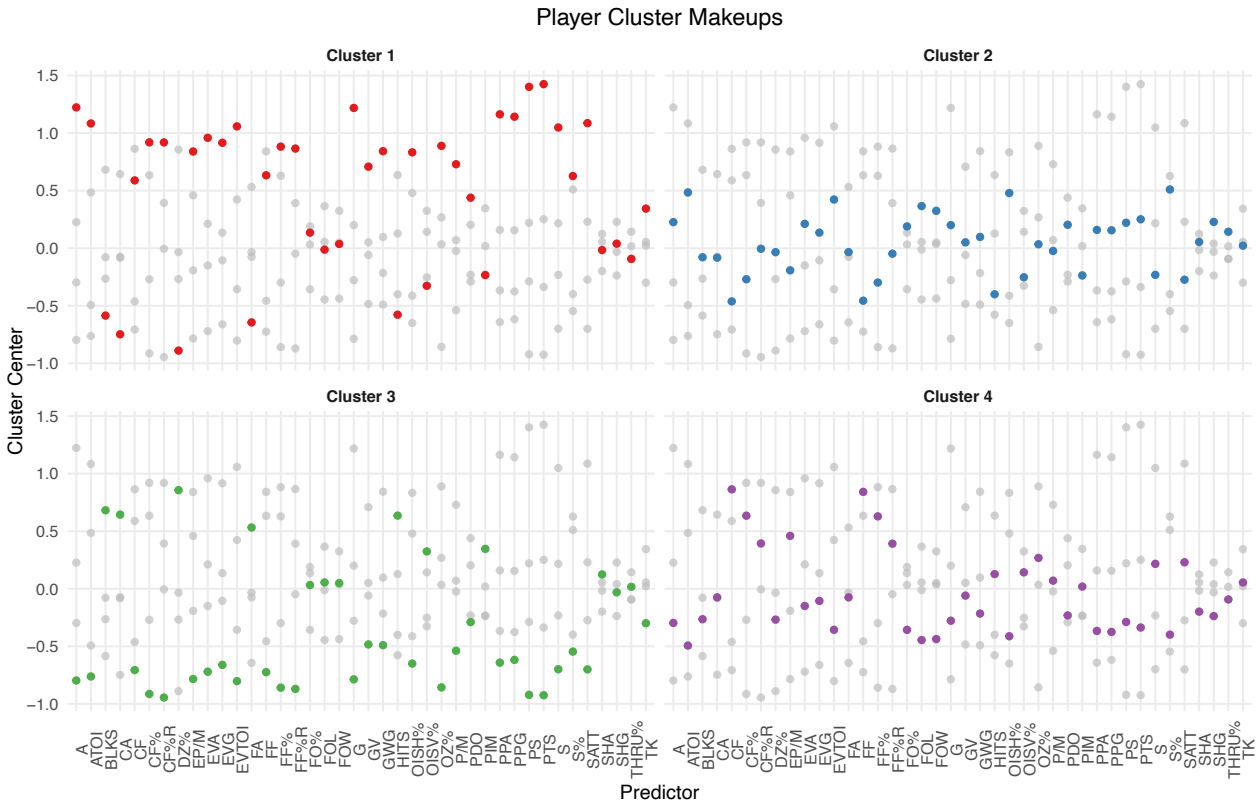


Figure 4. Forward cluster centers by scaled predictor variables for all forwards in the 2020-2021, 2021-2022, and 2022-2023 NHL seasons.

A very important note is that the mean values plotted are of the standardized data that were fed to the principal component analysis, which is why the variables are on the same scale. With that in mind, the observer should focus only on the relationship between the centers, not the values of the centers. For example, from the above plot it can be seen that Cluster 1 appears to be at the top of variables, while Cluster 3 appears to be at the bottom. Clusters 2 and 4 seem to be sitting in the middle ground. Abbreviations for the predictor variables were used to conserve space in the visualization, a key of said abbreviations and their respective variables can be found in Appendix A. Should the reader like a closer view of this visualization, it can be found in Appendix C. From these relationships and additional analysis, the player clusters were named and described. Such descriptions as well as players who fall in the clusters can be found in Table 2.

Table 2. Forward player cluster names, characteristics, and example players.

Cluster	Name	Characteristics	Notable Players
1	Franchise Player	Highest Scoring (G, A, PTS); Most Giveaways & Takeaways; Highest PlusMinus; Most Game-Winning Goals	Connor McDavid David Pastranak Dylan Larkin
2	Clutch Shooter	Highest Faceoff Percentage; Highest Thru Percentage; Most Short-Handed Goals; Second-Most Short-Handed Assists	Bo Horvat Dylan Cozens Zach Hyman
3	Defensive Hitter	Most Hits & Blocks; Highest Corsi & Fenwick Against; Highest On-Ice Save Percentage; Least Giveaways; Highest Penalty Minutes	Frank Vatrano Michael Rasmussen Adam Lowry Jeff Carter
4	Shooting Playmaker	Highest Corsi For; Highest Fenwick For; Second-Highest PlusMinus; Second-Most Shots	Nino Niederreiter Evgenii Dadanov Oliver Bjorkstrand

The first cluster of players are those typically considered “superstars.” They are named franchise players as organization’s typically acquire them in a first-round pick and build their team around them for the remainder of the player’s career. Franchise players are the highest scoring with the most goals and assists as well as the highest plusminus. With this additional possession time, franchise players do have the highest number of giveaways of the clusters, but easily make up for them by maintaining the highest number of takeaways.

The second cluster of players, the clutch shooters, are there when you need them most. These players are the most reliable with the highest number of faceoff wins and highest faceoff percentage. While they don’t take as many shots as other clusters, they maintain a shooting percentage just below the franchise players and the highest thru percentage. Clutch shooters

support their teams in the most crucial moments by having the most short-handed goals and the second most short-handed assists.

The defensive hitters, or third cluster of players, will put their body on the line to prevent an opposing goal. They have the highest number of blocks and the highest number of hits compared to the other three clusters. They face the most shots as they have the highest corsi against and fenwick against. While they may not contribute as much offensively, they ensure they will not feed the opposing defense by maintaining the lowest number of giveaways. Defensive hitters also aid their goalies as they are observed to have the highest team on-ice save percentage.

The final cluster of players, the shooting playmakers, do everything they can to get the puck to the net. They have the second highest number of shots as well as the highest corsi for and the highest fenwick for. Although these players may not be scoring the goals themselves, they are contributing on the ice when they happen as they have the second highest plusminus behind the franchise players. An additional aspect of their playmaking is their ability to take away the puck with the second highest number of takeaways. Shooting playmakers can be thought of as quiet contributors who use puck movement as their weapon of choice.

To examine if position had an impact on clustering, positional makeup was calculated for each cluster and can be found in Table 3.

Table 3. Positional makeup and totals of forward clusters.

Cluster	C	F	LW	RW	W	Total
1	156	8	101	68	3	336
2	254	11	84	66	0	415
3	264	25	114	83	2	488
4	158	12	107	113	2	392

Upon creating the above data table, a Pearson's Chi-squared test was performed on the two variables – cluster and position. The test returned a p-value of $p < 0.05$, suggesting that cluster and position are not independent. Thus, position did have an impact on clustering. In Table 3 it can be observed that centers and left wings make up the majority of all clusters with the most right wings being assigned to Cluster 4.

In comparing the results of the forward cluster analysis to those of the relevant literature, there are some similarities and differences observed. Schulte at al. (2017) also found all forward positions to be distributed amongst clusters with the majority of three of their four clusters consisting of forwards and right wings. A difference in Schulte et al.'s (2017) work was distribution of player numbers. They found three clusters to be relatively the same size with the fourth exceptionally smaller. Stimson's (2017) clusters were distributed more like those in this study, but two were roughly half the size of the others. Given both studies had smaller sample sizes and used different player metrics, it's feasible that their clusters were distributed differently. A final similarity with Stimson (2017) was that certain clusters were found to be stronger at scoring while others were not. While this conclusion could be made logically, agreeance of this study's results with previous work supports the notion that clusters created are accurate to playing styles. With the completion of cluster analysis for forwards, the same process was repeated for defensemen clusters.

Defensemen Clusters

As mentioned previously, defensemen were divided into 3 clusters rather than 4. Again, in examining the clusters the focus will begin on general patterns then narrow down to individual characteristics of clusters. Figure 5 illustrates cluster center locations by predictor variable.

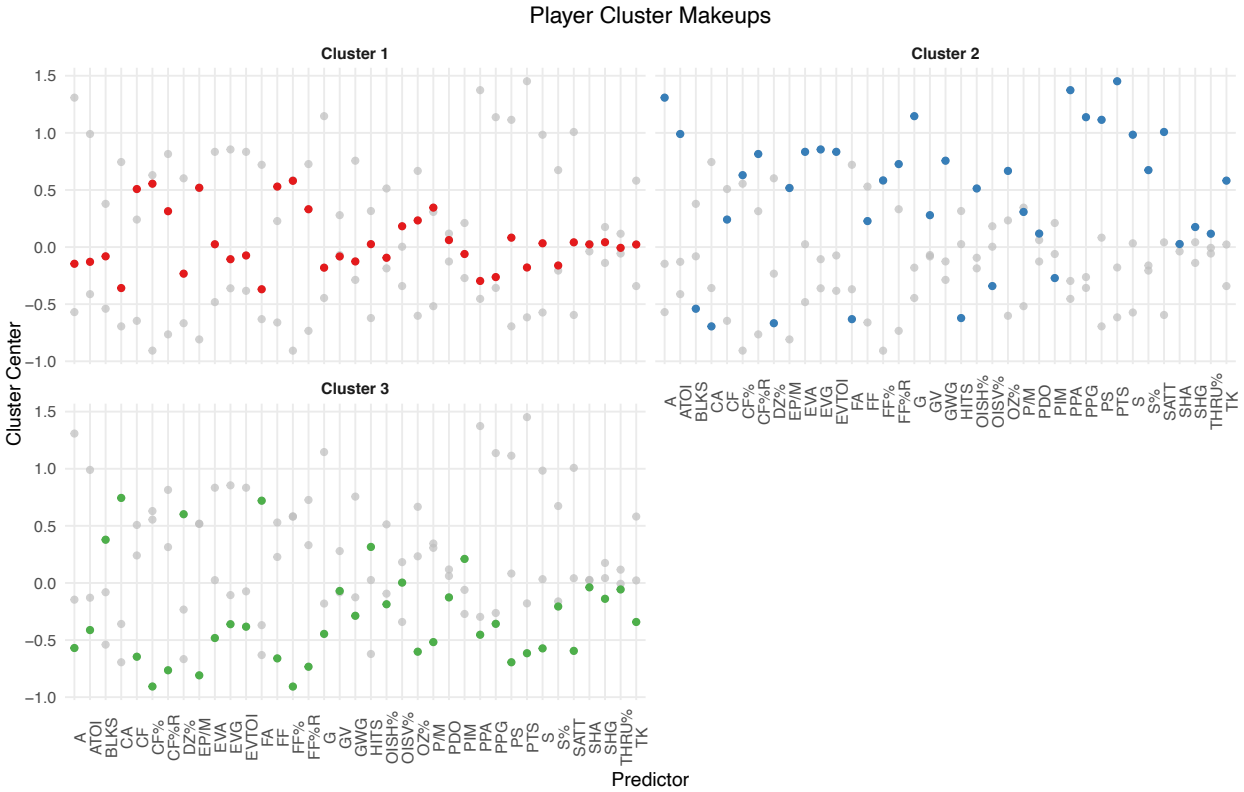


Figure 5. Defensemen cluster centers by scaled predictor variables for all forwards in the 2020-2021, 2021-2022, and 2022-2023 NHL seasons.

As with Figure 4, note that the mean values presented are on scaled data and thus the relationships between the values rather than the values themselves should be considered. A larger version of this graphic for closer examination can also be found in Appendix C. Figure 5 highlights that Cluster 2 has the highest values of most variables, Cluster 3 has the lowest, and Cluster 1 is in the middle. A deeper description of each cluster and exemplary players can be found in Table 4.

Table 4. Defensemen player cluster names, characteristics, and example players.

Cluster	Name	Characteristics	Notable Players
1	Hybrid Shooter	Highest Corsi For & Fenwick For; Highest PlusMinus; Highest On-Ice Save Percentage; Least Giveaways	Neal Pionk Matt Roy Jake Walman

2	Point Producer	Highest Scoring (G, A, PTS); Most Giveaways & Takeaways; Highest Average Time On Ice; Most Shots on Goal; Highest Shooting Percentage	Cale Makar Erik Karlsson Roman Josi
3	True Defender	Most Blocks; Most Hits; Highest Corsi & Fenwick Against; Highest Penalty Minutes	Esa Lindell Ryan McDonagh Nick Leddy

The first defensemen cluster, hybrid shooters, walk the line between the other two clusters which are more offensive and more defensive in nature. As for offense, hybrid shooters have both the highest corsi for and highest fenwick for, meaning they work hard to get the puck to the net. They provide key support to those who do score as they have the highest plusminus of the three clusters. In defensive terms, they prevent pucks from getting in net by having the highest team on-ice save percentage and the least giveaways. Hybrid shooters contribute to plays with puck movement while ensuring their own goal is untouched.

The point producers, or second cluster, are the defensive "superstars." As with the franchise player forwards, teams are built around point producers, and they typically stay with one team for their entire career. Point producers are the highest scoring cluster while also having the most shots and highest shooting percentage. They have the highest number of giveaways but also have the highest number of takeaways, both likely due to the fact that they have the highest average time on ice. Point producers can be thought of as more forward-thinking defensemen.

The final defensemen cluster are true defenders. As their name suggests, true defenders contribute to their teams via defensive measures rather than points. They have the highest number of blocks and hits, with the hits translating into them having the highest penalty minutes. They also face the most shots from the opposing team with the highest corsi against and highest fenwick against. True defenders are focused on stopping the opposing offensive in its tracks, even if means picking up a penalty. The resulting number of players in each defensive cluster can be found in Table 5.

Table 5. Number of players per defensive player cluster.

Cluster Name	Number of Players
Hybrid Shooter	343
Point Producer	186
True Defender	339

Due to the lack of differentiated positions in the defensemen data set, positional makeup of clusters was not calculated for defensemen. Non-differentiated defensemen positions were the only similarity between Schulte et al. (2017) and this study. A major difference with both Schulte et al. (2017) and Stimson (2017) was the number of defensemen clusters. As mentioned previously, the two studies created 4 clusters while this study found 3 to be optimal after rigorous testing (Table 1). In Table 5 it can be observed that the highest scoring player cluster was found to have the least number of players, the same was found to be true for Stimson (2017). Once again, similar distribution of player clusters by previous work most similar to this project is

taken as support for resulting clusters. After clustering was performed on all data, Detroit's current roster was examined.

Detroit's Roster

With an understanding of the strengths and weaknesses of each cluster, considering Detroit's Roster with player cluster assignments provides unique insight into the composition of the team. First, all forwards who play with Detroit and were above the games played cutoff of 15 games were examined. The results can be found in Table 6.

Table 6. Detroit Red Wings 2022-2023 forwards roster with stats and cluster assignments.

Player	Age	Pos	GP	G	A	+/-	PIM	S	S%	ATOI	Cluster	Cluster Name
Jonatan Berggren	22	RW	67	15	13	-14	16	98	15.3	13.47	2	Clutch Shooter
Alex Chiasson	32	RW	20	6	3	-8	6	25	24	12.07	2	Clutch Shooter
Andrew Copp	28	C	82	9	33	2	25	120	7.5	18.15	2	Clutch Shooter
Austin Czarnik	30	C	29	3	2	-4	8	27	11.1	11.17	3	Defensive Hitter
Adam Erne	27	LW	61	8	10	-12	21	55	14.5	13.38	3	Defensive Hitter
Robby Fabbri	27	C	28	7	9	-1	22	35	20	16.00	2	Clutch Shooter
Dominik Kubalík	27	LW	81	20	25	-15	24	174	11.5	14.92	2	Clutch Shooter
Dylan Larkin	26	C	80	32	47	-7	45	244	13.1	19.55	1	Franchise Player
Matt Luff	25	RW	19	2	2	-4	0	20	10	9.88	3	Defensive Hitter
David Perron	34	LW	82	24	32	-7	52	195	12.3	16.92	1	Franchise Player
Michael Rasmussen	23	C	56	10	19	2	43	88	11.4	15.08	2	Clutch Shooter
Lucas Raymond	20	LW	74	17	28	-17	24	134	12.7	17.38	2	Clutch Shooter
Elmer Söderblom	21	RW	21	5	3	0	8	30	16.7	12.07	4	Shooting Playmaker
Pius Suter	26	F	79	14	10	-3	6	106	13.2	14.07	2	Clutch Shooter
Joe Veleno	23	C	81	9	11	-12	30	85	10.6	12.78	3	Defensive Hitter
Filip Zadina	23	RW	30	3	4	-5	10	51	5.9	13.10	4	Shooting Playmaker

Of the 16 forwards, 50% are clutch shooters, 25% are defensive hitters, 12.5% are franchise players, and 12.5% are shooting playmakers. Given that clutch shooters and franchise players are the top two clusters in scoring, the fact that these types of players make up a combined 62.5% of the current roster suggests the offensive portion of the team is solid. The biggest surprise of the forwards analysis was David Perron being classified as a franchise player despite this season being his first with the team. It also should be noted that the team has an excellent set of young players – Berggren, Raymond, and Rasmussen - who are nearly at the franchise player level in the beginning on their careers. Next, all defensemen who played 13 games or more with Detroit were examined and results can be found in Table 7.

Table 7. Detroit Red Wings 2022-2023 defensemen roster with stats and cluster assignments.

Player	Age	Pos	GP	G	A	+/-	PIM	S	S%	ATOI	Cluster	Cluster Name
Ben Chiarot	31	D	76	5	14	-31	51	110	4.5	20.62	3	True Defender
Robert Hägg	27	D	38	2	5	-5	26	40	5	15.52	3	True Defender
Gustav Lindström	24	D	36	1	7	-16	20	18	5.6	14.17	3	True Defender
Olli Määttä	28	D	78	6	17	-9	14	62	9.7	18.70	3	True Defender

Jordan Oesterle	30	D	52	2	9	-9	19	54	3.7	15.65	3	True Defender
Moritz Seider	21	D	82	5	37	-11	40	140	3.6	23.15	1	Hybrid Shooter
Jake Walman	26	D	63	9	9	10	45	140	6.4	19.72	1	Hybrid Shooter

Defensively, the Red Wings are in a precarious position. With 71.4% of the defensemen being in the lowest scoring cluster, 28.6% in the second lowest scoring cluster, and none in the highest scoring cluster, it's clear there is a hole in Detroit's defense. Perhaps the most shocking revelation to come from the defensemen clustering analysis is the placement of Moritz Seider in the hybrid shooter cluster. Likely attributable to the infamous "sophomore slump", Seider struggled to find scoring success, at least in goals, this past season and went from the high-scoring point producer cluster in his rookie season to the middle-ground hybrid shooter cluster. One glimmer of hope in the Wings' defense is found in Jake Walman. With the most goals and highest plusminus of the defensemen, Walman appears to be working his way up to the point producer cluster as he continually grows stronger with Detroit. With the current state of the Detroit's offensive and defensive rosters in mind, an optimization was performed to identify roster moves that will improve the overall success of the team.

Roster Optimization

Of the previous studies reviewed, only Stimson (2017) considered optimization. However, Stimson focused on optimizing player lines, rather than whole-team composition. A team-based approach provides the unique benefit of identifying the ideal number of player cluster types. In addition, Stimson did not utilize constrained optimization, rather he just compared all possible options of line combinations. Using constrained optimization ensures that all possible options are considered and the ideal selected. Not to mention, it's customizability allows for easy manipulation of constraints and objectives at management's request. The power of constrained optimization can be understood by the optimization of Detroit's roster.

With more players considered in the previous section than would be allowed on a roster, a hypothetical 21-man roster was created with 14 forwards and 7 defensemen. To determine what forwards were included in the roster, the points per games played statistic was used. This decision was made as clustering was performed on standardized statistics, so it was appropriate that roster creation was as well. Additionally, players with too few games played were already filtered out of the data set which will prevent too harsh of skewing in the statistic itself. The two forwards with the lowest points per games played and were thus not included in the hypothetical roster were Matt Luff and Austin Czarnik.

The hypothetical Detroit roster's total points per games played and total cap hit were used as a baseline for comparison of alternative team composition options, as can be found in Table 8.

Table 8. Original team roster and roster optimization player cluster makeups, average points per games played, and cap hit.

Cluster Name	Original Roster	Option 1	Option 2	Option 3	Option 4
Franchise Player	2	3	2	3	2
Defensive Hitter	2	2	3	2	1
Clutch Shooter	8	8	6	7	8
Shooting Playmaker	2	1	3	2	3
True Defender	5	3	4	2	3

Point Producer	0	2	1	2	1
Hybrid Shooter	2	2	2	3	3
Total Points/Games Played	8.72	10.88	10.24	10.64	10.62
Cap Hit	$ 47,895,833	$ 51,920,833	$ 53,845,833	$ 56,995,833	$ 54,701,666

It should be noted that each optimization option was constrained to retain 17 of the 21, or approximately 81%, of the original roster to prevent too drastic of changes. The optimizations were also constrained to include 14 forwards and 7 defensemen as well as the unique number of player cluster types specified in their respective columns of Table 8. Calculations of total points per games played and total cap hit were made assuming that players perform and are paid at a similar rate as the 2022-2023 season, thus totals are estimations for next season.

Each option was crafted by making a small adjustment to the dispersion of forward and defensemen player cluster types. Option 1, highlighted in Table 8, was found to improve total points per games played the most while also increasing cap hit the least. In fact, Option 1 only increases cap hit by 8.4% but increases total points per games played by 24.8%, making it well worth the tradeoff. The suggested changes in Option 1 are eliminating one shooting playmaker in exchange for a franchise player and eliminating two true defenders in exchange for two point producers. The shooting playmaker to be dropped was Filip Zadina in exchange for the New Jersey Devils' Jesper Bratt. The two true defenders were Robert Hägg and Jordan Oesterle in exchange for Vince Dunn of the Seattle Kraken and Erik Gustafsson of the Toronto Maple Leafs. Given the optimization had the power to replace four players, it also replaced one clutch shooter, Pius Suter, for another, David Krejčí, of the Boston Bruins. The entire optimized roster and accompanying general stats can be found in Appendix D.

Realistically, it's unlikely that Krejčí would finish out his career with the Wings being a lifelong Bruins player and the current alternate captain of the team. However, this does not discount the optimization's recommendations – the most important insights gained are which Red Wings were dropped, and the cluster type of the players picked up in their stead. The recommended players can easily be interchanged by tweaking the optimization constraints to meet management's needs. For example, should an age limitation of equal to or less than 30 years old be placed on the available players, the optimized roster remains the same besides replacing Krejčí with Matias Maccelli of the Arizona Coyotes. The total points per games played is reduced by only 0.03 while the total cap hit is reduced by $146,667, creating a more realistic and cost-efficient roster composition. The age-limited optimized roster and accompanying general stats can be found in Appendix E.

The results of the optimized roster in combination with the cluster analysis and examination of Detroit's roster were factored into compiling the final recommendations for the team.

Conclusion

Key Takeaways

Prior to the discussion of recommendations, consider what has been learned from each aspect of the analysis:

- Forward Clusters – Forwards fall into four types of player clusters: franchise players, clutch shooters, defensive hitters, and shooting playmakers. Examples of each on the Red Wings are Dylan Larkin, Dominik Kubalík, Joe Veleno, and Filip Zadina, respectively.
- Defensemen Clusters – Defensemen fall into three types of player clusters: hybrid shooters, point producers, and true defensemen. Detroit currently has no point producers, but examples of a hybrid shooter and true defensemen are Jake Walman and Ben Chiarot, respectively.
- Detroit's Roster – Offensively, the team has a solid combination of forward player cluster types with 50% clutch shooters, 25% defensive hitters, 12.5% franchise players, and 12.5% shooting playmakers. Defensively, the team is lacking with 71.4% true defenders, 28.6% hybrid shooters, and no players of the highest scoring cluster, point producers.
- Roster Optimization – The most efficient way to increase points per games played while keeping cap hit low is to increase the number of franchise players from 2 to 1, decrease the number of shooting playmakers from 2 to 1, decrease the number of true defenders from 5 to 3, and increase the number of point producers from 0 to 2. Exchanging one clutch shooter for another was also recommended.

Recommendations

From the observed results of the analysis, the following recommendations are suggested for the Detroit Red Wings: (1) Look for young franchise players and clutch shooters, (2) rebuild the defense with a focus on point producers, and (3) consider trading away Filip Zadina, Pius Suter, and Jordan Oesterle. The first recommendation comes from the optimization discovering Matias Maccelli of the Arizona Coyotes as a replacement clutch shooter. Now that players have been classified into clusters, management can find hidden gems to acquire. Let this be evidenced by the fact that last season Maccelli scored more goals than 10 forwards on Detroit's roster, and he's only a clutch shooter. Looking for franchise players has the potential for even higher point contribution. One example of a potentially underrated, young franchise player that could be acquired is Max Comtois of the Anaheim Ducks. Last season he scored more goals than 12 forwards on Detroit's roster. With the clustering algorithm's ability to identify play type from a combination of many variables, it can provide scouting and management with the ability to see a player's true potential in their playing style.

The second recommendation stems from the defensive roster analysis of Detroit and it's glaring shortcomings. As the roster optimization discovered, shifting the majority of the roster from true defenders to both point producers and hybrid shooters will translate to higher scoring. However, acquiring new players may not be required to achieve a more balanced roster. Both Moritz Seider and Jake Walman are on the cusp of the point producer cluster and with rigorous training this offseason, they have the potential to cross the threshold next season. Should they increase point output, then it's also recommended that prospects from lower league be brought up and played to identify playing style. This will allow the team to get a better idea of the type of player their prospects are. The bottom line is that Detroit's defense needs to improve next season. Whether this is done internally or externally is at management's discretion. If the decision is made to look externally, as with identifying young franchise players, the same can be done for point producers.

The final recommendation is to consider trading away Filip Zadina, Pius Suter, and Jordan Oesterle. In every iteration of the roster optimization, it was recommended that these three players be replaced. Zadina has been with the team for three seasons and has

underperformed compared to the massive expectations that the team had for him when using their first-round draft pick on him in 2018. As such, he's had ample time to improve and become comfortable in the NHL, yet he can't rise above the shooting playmaker cluster. Pius Suter is another example of a player who's had ample time with the Red Wings, yet his point production continues to decrease with each season. Although he's considered a clutch shooter now, he'll likely fall to shooting playmaker in the near future. Finally, since joining the Red Wings from the Coyotes in the 2021-2022 season, Oesterle has yet to find his stride. All his recent years of play he has been classified as a true defender, showing little growth in his stats. Given he's 30 years old, he may be coming out of his prime. Overall, trading away these players will allow Detroit to build a stronger team. With these recommendations being informed by the results of the analyses performed, several limitations of the analyses themselves should be addressed.

Limitations

As was alluded to in the data cleaning and filtering section, there were some problems with the quality of the player performance data fed to the clustering model. The issues appeared to stem from data aggregation of players who were on more than one team as well as changing the format of recorded stats year-to-year. While this was inconvenient, relatively few players were impacted. With that being said, when using a publicly available data source like hockeyreference.com, it's impossible to validate the accuracy of all entries. After all, results are only as good as the data they were built on.

Another limitation of this study was selecting the best player performance metrics to use in the clustering algorithm and objective function of the constrained optimization. The truth is there is more to a player than the general recorded stats of a game which often don't encapsulate qualities like speed, location on the ice, or quality of shots. As such, part of the goal of this project was to find meaning in the stats that are available to everyone, which means accepting they are less descriptive of a player as a whole.

The final two limitations are related to principal component analysis. First, the number of principal components selected to be input into the clustering algorithm has a significant impact on the results. While testing was performed to this parameter, additional confirmation that the optimal number was chosen should be sought out. Second, outstanding players like Connor McDavid perform so well that their stats are considered outliers compared to the rest of the players in the NHL. Outliers, like McDavid's stats, have been proven to negatively impact both the scaling of data and PCA (Sapra 2010). While the severity of the impact from the outliers in this data set is unknown, it should be considered when interpreting results. The identified limitations of this analysis will be used as a guide for future research.

Future Research

To address data quality issues, the next iteration of this study will use data sourced directly from the Detroit Red Wings and other teams in the league. By obtaining official data, it will be of the highest quality and verified for accuracy. To initiate this process, contact will be made with data managers, or other relevant personnel, on each team via email. From there, discussion can begin over the best method of data sharing.

Improving data quality will also help to address selection of the best player performance metrics by ensuring metrics are accurate to the player. In selecting the best metrics to input into the clustering algorithm, future research will compare and contrast clustering results on different metrics including raw statistics, statistics standard by total minutes played, statistics standardized

by games played, and statistics standardized by average time on ice. In selecting the best metric for the objective function in constrained optimization, future studies will be performed to identify the best, single player performance statistic. Research will follow the methodology used by Luszczyszyn (2016) in combining several basic stats to create one representative of the player's contributions per game. Upcoming actions to be taken for this research are the creation of the aformentioned sets of variables and running them through the clustering code as well as beginning literature review for relevant studies looking at creating a single player performance metric.

Technical solutions will be utilized to identify the optimal number of principal components and handle outlier impact on scaling and PCA. To identify the optimal number of principal components, a permutation test will be run as outlined by Toledo Jr. (2022). The test will randomly sample each column of the data set, effectively decorrelating the data, and run PCA on the sample. The idea is that if a principal component is relevant, the explained variance of the principal component on the original data set should be greater than that on the permutated data set. The next step to initiate this process will simply be to input the relevant code into the current R document. As for handling outlier impact, robust scaling and robust PCA are specialized algorithms that are resistant to outliers. Robust PCA has been proven to reduce data sets more efficiently than classical PCA in data sets with outliers (Sapra 2010). For this reason, robust scaling and robust PCA will be implemented in future research for this study. To do so, the first step will require downloading the relevant R libraries and performing research into implementing them in the current notebook.

An additional goal beyond addressing limitations in future research would be to expand the clustering algorithm and constrained optimization analysis to prospect data. By classifying playing styles of prospects, the team would have a better understanding of what to expect from a player when their pulled up to the NHL. The hope is that the scouting team and management could use cluster classifications as a tool in evaluating players. To begin this process, data from all leagues in which Red Wings prospects participate will need to be acquired. This may be a lengthy process as prospects are typically dispersed all over the world. An additional data set that may provide new insights in the analysis is the annual NHL Scouting Combine results. Such results will be far easier to obtain as they are posted on a singular website. With the next steps to improve this study and expand upon its original findings in place, work can begin to help Detroit secure their next Stanley Cup.

Appendix A: Data Dictionary

Variable Name	Description	Inclusion in Forward Clustering	Inclusion in Defensemen Clustering	Abbreviation in Visualization
player	player name	No	No	
age	player age	No	No	
tm	team	No	No	
pos	player position	No	No	
gen_pos	generalized player position (only F and D)	No	No	
gp	games played	No	No	
g	goals	No	No	
stand_g	goals / time on ice (minutes)	Yes	Yes	G
a	assists	No	No	
stand_a	assists / time on ice (minutes)	Yes	Yes	A
pts	points	No	No	
stand_pts	points / time on ice (minutes)	Yes	Yes	PTS
plusminus	plus minus; sum of goals scored and goals given up while player is on the ice	No	No	
stand_plusminus	plus minus / time on ice (minutes)	Yes	Yes	P/M
pim	penalties in minutes	No	No	
stand_pim	penalties in minutes / time on ice (minutes)	Yes	Yes	PIM
ps	point shares; an estimate of the number of points contributed by a player	No	No	
stand_ps	point shares / time on ice (minutes)	Yes	Yes	PS
ev_g	even strength goals	No	No	
stand_ev_g	even strength goals / time on ice (minutes)	Yes	Yes	EVG
pp_g	power play goals	No	No	
stand_pp_g	power play goals / time on ice (minutes)	Yes	Yes	PPG
sh_g	short-handed goals	No	No	
stand_sh_g	short-handed goals / time on ice (minutes)	Yes	Yes	SHG
gw_g	game-winning goals	No	No	
stand_gw_g	game-winning goals / time on ice (minutes)	Yes	Yes	GWG
ev_a	even strength assists	No	No	
stand_ev_a	even strength assists / time on ice (minutes)	Yes	Yes	EVA
pp_a	power play assists	No	No	

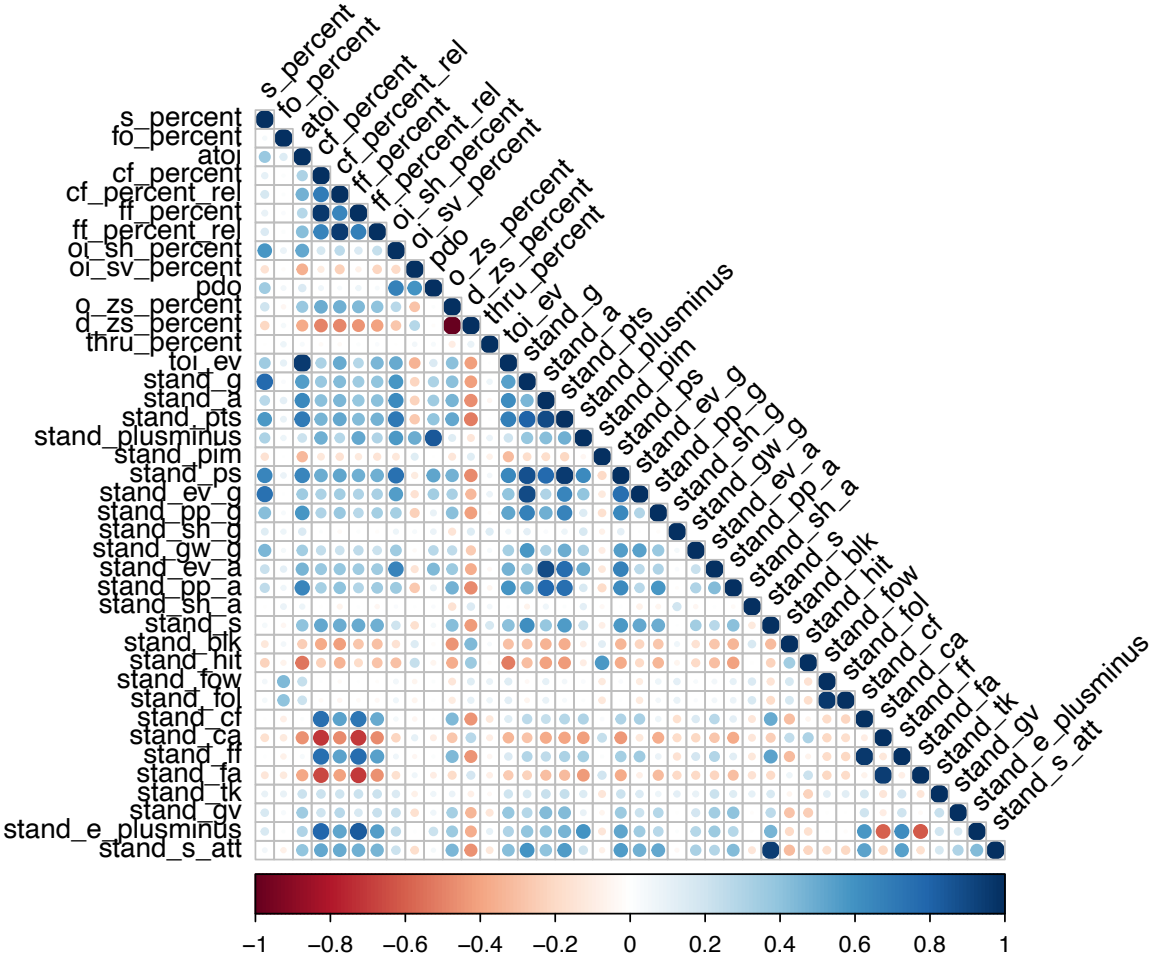
stand_pp_a	power play assists / time on ice (minutes)	Yes	Yes	PPA
sh_a	short-handed assists	No	No	
stand_sh_a	short-handed assists / time on ice (minutes)	Yes	Yes	SHA
s	shots on goal	No	No	
stand_s	shots on goal / time on ice (minutes)	Yes	Yes	S
s_percent	shooting percentage	Yes	Yes	S%
toi	time on ice (minutes)	No	No	
atoi	average time on ice	Yes	Yes	ATOI
blk	blocks	No	No	
stand_blk	blocks / time on ice (minutes)	Yes	Yes	BLKS
hit	hits	No	No	
stand_hit	hits / time on ice (minutes)	Yes	Yes	HITS
fow	faceoff wins	No	No	
stand_fow	faceoff wins / time on ice (minutes)	Yes	No	FOW
fol	faceoff losses	No	No	
stand_fol	faceoff losses / time on ice (minutes)	Yes	No	FOL
fo_percent	faceoff win percentage	Yes	No	FO%
season	NHL season	No	No	
cf	corsi for at even strength (shots + blocks + misses)	No	No	
stand_cf	corsi for at even strength (shots + blocks + misses) / time on ice (minutes)	Yes	Yes	CF
ca	corsi against at even strength (shots + blocks + misses)	No	No	
stand_ca	corsi against at even strength (shots + blocks + misses) / time on ice (minutes)	Yes	Yes	CA
cf_percent	corsi for percentage at even strength (CF / CF + CA) - Above 50% means the team was controlling the puck more often than not with this player on the ice in this situation	Yes	Yes	CF%
cf_percent_rel	realtive corsi for percentage at even strength (CF% - Cfoff%) On-ice corsi for percentage minus off-ice corsi for percentage	Yes	Yes	CF%R
ff	fenwick for at even strength (shots + misses)	No	No	

stand_ff	fenwick for at even strength (shots + misses) / time on ice (minutes)	Yes	Yes	FF
fa	fenwick against at even strength (shots + misses)	No	No	
stand_fa	fenwick against at even strength (shots + misses) / time on ice (minutes)	Yes	Yes	FA
ff_percent	fenwick for percentage at even strength (FF / FF + FA) - Above 50% means the team was controlling the puck more often than not with this player on the ice in this situation	Yes	Yes	FF%
ff_percent_rel	realtive fenwick for percentage at even strength (FF% - FFoff%) On-ice fenwick for percentage minus off-ice fenwick for percentage	Yes	Yes	FF%R
oi_sh_percent	team on-ice shooting percentage - shooting percentage while this player/team was on the ice	Yes	Yes	OISH%
oi_sv_percent	team on-ice save percentage - save perentage while this player/team was on the ice	Yes	Yes	OISV%
pdo	pdo - shooting percentage + save percentage	Yes	Yes	PDO
o_zs_percent	offensive zone start percentage - offensive zone faceoffs / (offensive zone faceoffs + defensive zone faceoffs) that took place while on the ice	Yes	Yes	OZ%
d_zs_percent	defensive zone start percentage - defensive zone faceoffs / (offensive zone faceoffs + defensive zone faceoffs) that took place while on the ice	Yes	Yes	DZ%
tk	takeways	No	No	
stand_tk	takeways / time on ice (minutes)	Yes	Yes	TK
gv	giveaways	No	No	
stand_gv	giveaways / time on ice (minutes)	Yes	Yes	GV
e_plusminus	expected plus/minus - given where the shots came from, for and against, while this player was on the ice at even strength. It's based on where the shots are coming from, compared to the	No	No	EP/M

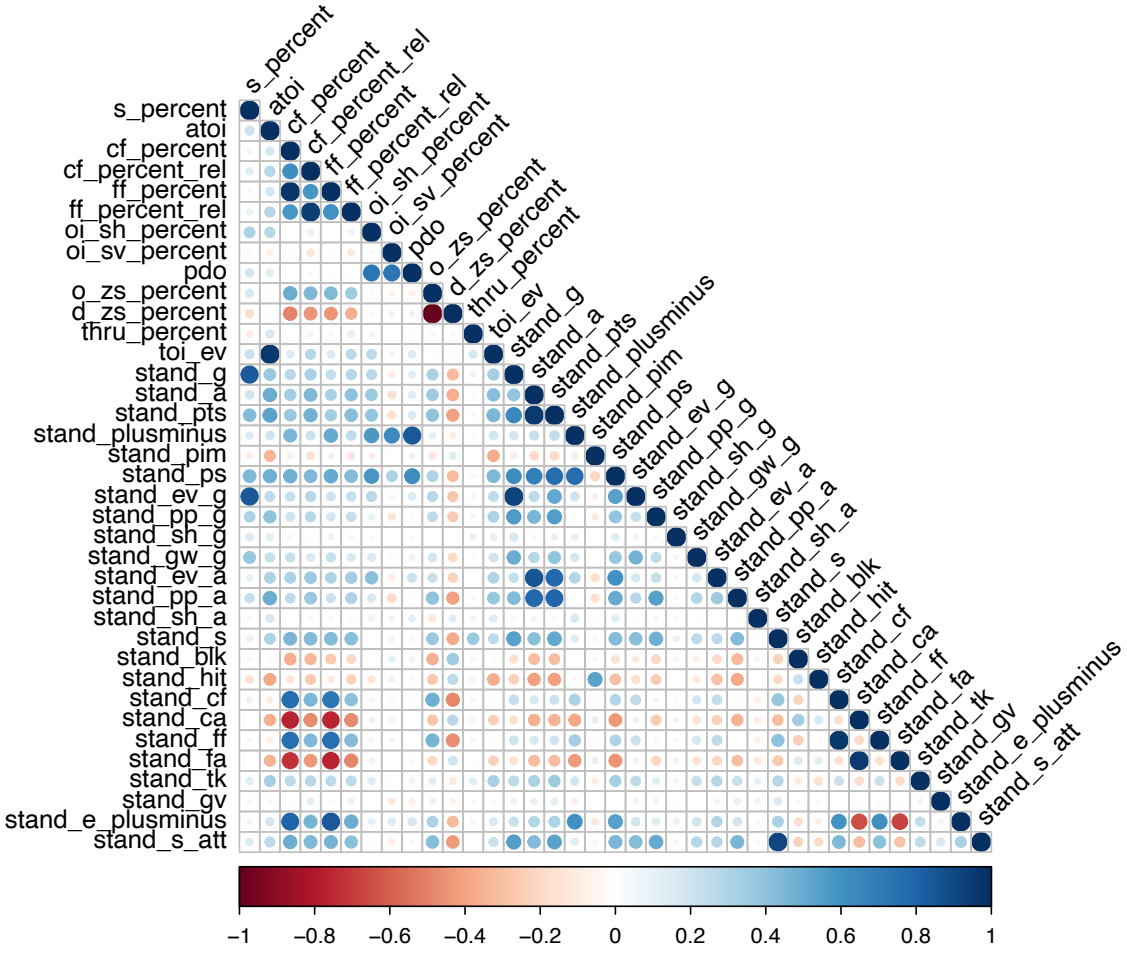
	league-wide shooting percentage for shot location			
stand_e_plusminus	expected plus/minus / time on ice (minutes)	Yes	Yes	
s_att	total shots attempted in all situations	No	No	
stand_s_att	total shots attempted in all situations / time on ice (minutes)	Yes	Yes	SATT
thru_percent	percentage of shots taken that go on net	Yes	Yes	THRU%
toi_60	toi/60 in all situations - time on ice per 60 minutes	No	No	
toi_ev	toi/60 at even strength - time on ice per 60 minutes	No	No	EVTOI

Appendix B: Correlation Plots

Forwards

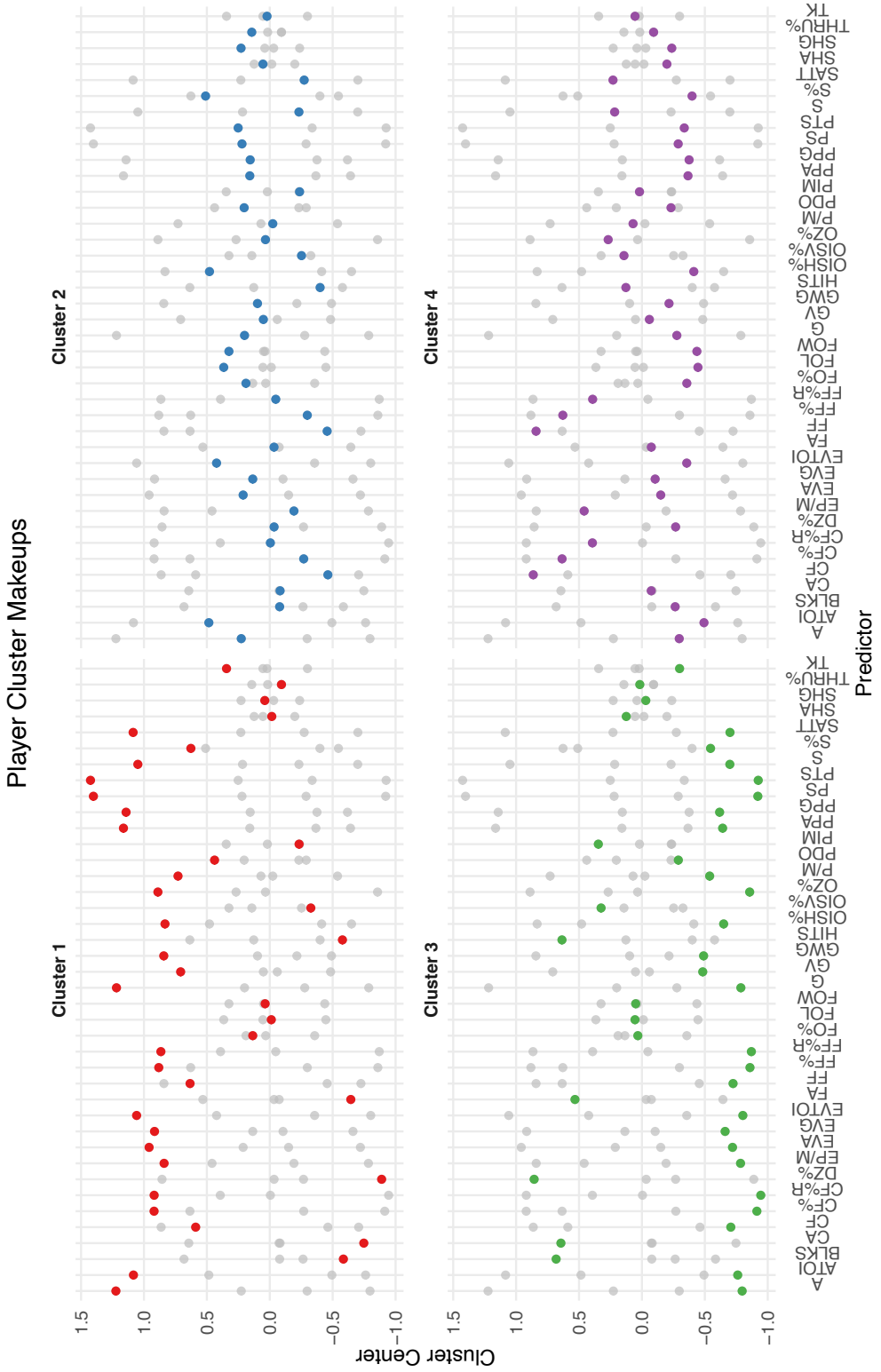


Defensemen

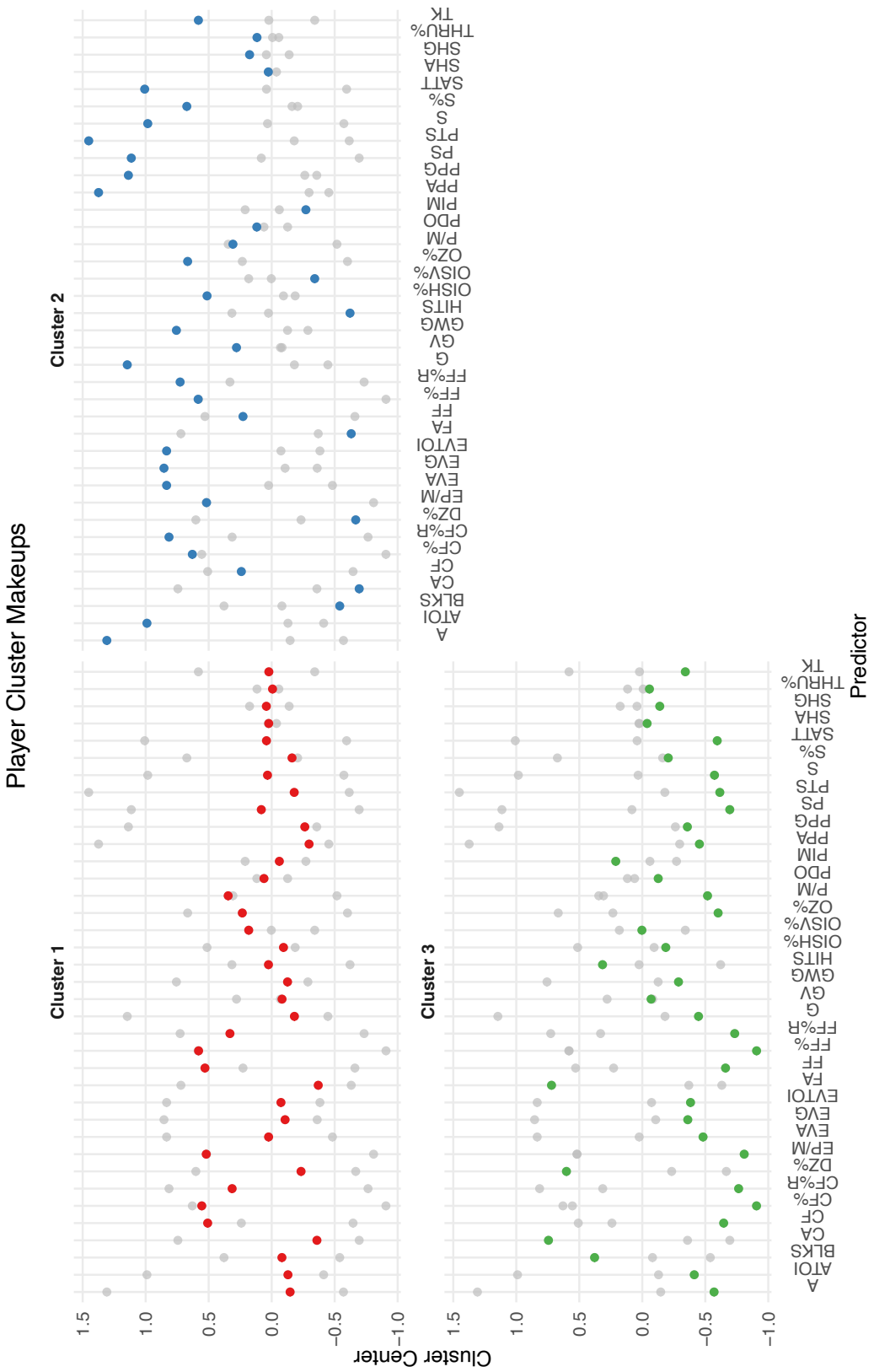

s_percent
atoi
cf_percent
cf_percent_rel
ff_percent
ff_percent_rel
oi_sh_percent
oi_sv_percent
pdo
o_zs_percent
d_zs_percent
thru_percent
toi_ev
stand_g
stand_a
stand_pts
stand_plusminus
stand_pim
stand_ps
stand_ev_g
stand_pp_g
stand_sh_g
stand_gw_g
stand_ev_a
stand_pp_a
stand_sh_a
stand_s
stand_blk
stand_hit
stand_cf
stand_ca
stand_ff
stand_fa
stand_tk
stand_gv
stand_e_plusminus
stand_s_att
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1

Appendix C: Player Cluster Makeups

Forwards



Defensemen


Player Cluster Makeups
Cluster 1
Cluster 2
Cluster 3
Cluster Center
1.5
1.0
0.5
0.0
−0.5
−1.0
Predictor
A
ATOI
BLKS
CA
CF
CF%
CF%R
DZ%
EPM
EVA
EVG
EVTOI
FA
FF
FF%
FF%R
G
GV
GWG
HITS
OISH%
OISV%
OZ%
PM
PDO
PIM
PPA
PPG
PS
PTS
S
S%
SATT
SHA
SHG
THRU%
TK

Appendix D: Optimized Roster

Player	Team	Age	Pos	GP	G	A	P/GP	+/-	S	S%	Cluster
Dylan Larkin	DET	25	C	80	32	47	0.99	-7	244	0.13	Franchise Player
Andrew Copp	DET	27	C, LW, RW	82	9	33	0.51	2	120	0.08	Clutch Shooter
David Perron	DET	34	RW, LW	82	24	32	0.68	-7	195	0.12	Franchise Player
Robby Fabbri	DET	26	LW, C	28	7	9	0.57	-1	35	0.2	Clutch Shooter
Dominik Kubalík	DET	26	LW, RW	81	20	25	0.56	-15	174	0.11	Clutch Shooter
Adam Erne	DET	27	LW, RW	61	8	10	0.3	-12	55	0.15	Defensive Hitter
Michael Rasmussen	DET	23	C, LW	56	10	19	0.52	2	88	0.11	Clutch Shooter
Jonatan Berggren	DET	21	RW, LW	67	15	13	0.42	-14	98	0.15	Clutch Shooter
Lucas Raymond	DET	20	RW, LW	74	17	28	0.61	-17	134	0.13	Clutch Shooter
Joe Veleno	DET	22	C	81	9	11	0.25	-12	85	0.11	Defensive Hitter
Elmer Söderblom	DET	20	LW, C	21	5	3	0.38	0	30	0.17	Shooting Playmaker
Alex Chiasson	DET	31	RW	20	6	3	0.45	-8	25	0.24	Clutch Shooter
Ben Chiarot	DET	31	LD/RD	76	5	14	0.25	-31	110	0.05	True Defender
Olli Määttä	DET	27	LD	78	6	17	0.29	-9	62	0.1	True Defender
Jake Walman	DET	26	LD	63	9	9	0.29	10	140	0.06	Hybrid Shooter
Moritz Seider	DET	21	RD	82	5	37	0.51	-11	140	0.04	Hybrid Shooter
Gustav Lindström	DET	23	RD	36	1	7	0.22	-16	18	0.06	True Defender
Jesper Bratt	NJD	23	RW, LW	82	32	41	0.89	14	212	0.15	Franchise Player
David Krejčí	BOS	36	C	70	16	40	0.8	23	112	0.14	Clutch Shooter
Vince Dunn	SEA	25	LD/RD	81	14	50	0.79	28	149	0.09	Point Producer
Erik Gustafsson	TOR	30	LD	70	7	35	0.6	9	122	0.06	Point Producer

Appendix E: Age-Limited Optimized Roster

Player	Team	Age	Pos	GP	G	A	P/GP	+/-	S	S%	Cluster
Dylan Larkin	DET	25	C	80	32	47	0.99	-7	244	0.13	Franchise Player
Andrew Copp	DET	27	C, LW, RW	82	9	33	0.51	2	120	0.08	Clutch Shooter
David Perron	DET	34	RW, LW	82	24	32	0.68	-7	195	0.12	Franchise Player
Robby Fabbri	DET	26	LW, C	28	7	9	0.57	-1	35	0.2	Clutch Shooter
Dominik Kubalík	DET	26	LW, RW	81	20	25	0.56	-15	174	0.11	Clutch Shooter
Adam Erne	DET	27	LW, RW	61	8	10	0.3	-12	55	0.15	Defensive Hitter
Michael Rasmussen	DET	23	C, LW	56	10	19	0.52	2	88	0.11	Clutch Shooter
Jonatan Berggren	DET	21	RW, LW	67	15	13	0.42	-14	98	0.15	Clutch Shooter
Lucas Raymond	DET	20	RW, LW	74	17	28	0.61	-17	134	0.13	Clutch Shooter
Joe Veleno	DET	22	C	81	9	11	0.25	-12	85	0.11	Defensive Hitter
Elmer Söderblom	DET	20	LW, C	21	5	3	0.38	0	30	0.17	Shooting Playmaker
Alex Chiasson	DET	31	RW	20	6	3	0.45	-8	25	0.24	Clutch Shooter
Ben Chiarot	DET	31	LD/RD	76	5	14	0.25	-31	110	0.05	True Defender
Olli Määttä	DET	27	LD	78	6	17	0.29	-9	62	0.1	True Defender
Jake Walman	DET	26	LD	63	9	9	0.29	10	140	0.06	Hybrid Shooter
Moritz Seider	DET	21	RD	82	5	37	0.51	-11	140	0.04	Hybrid Shooter
Gustav Lindström	DET	23	RD	36	1	7	0.22	-16	18	0.06	True Defender
Jesper Bratt	NJD	23	RW, LW	82	32	41	0.89	14	212	0.15	Franchise Player
Matias Maccelli	ARI	21	LW	64	11	38	0.77	0	61	0.18	Clutch Shooter
Vince Dunn	SEA	25	LD/RD	81	14	50	0.79	28	149	0.09	Point Producer
Erik Gustafsson	TOR	30	LD	70	7	35	0.6	9	122	0.06	Point Producer

References

Batra, Ayush. 2022. "Generating NBA Archetypes Using K-Means Clustering." Best Ball Stats, July 23, 2022, https://bestballstats.com/2022/07/23/generating-nba-archetypes-using-k-means-clustering/#:~:text=The%20NBA%20commonly%20uses%20five,%2C%20power%20forward%2C%20and%20center.

Hussain, Haider. 2019. "Using K-Means Clustering Algorithm to Redefine NBA Positions and Explore Roster Construction." Towards Data Science, November 7, 2019. https://towardsdatascience.com/using-k-means-clustering-algorithm-to-redefine-nba-positions-and-explore-roster-construction-8cd0f9a96dbb.

Lähevirta, Elmeri. 2018. "On cluster structures of NHL players." Aalto University School of Science, January 15, 2018. https://sal.aalto.fi/publications/pdf-files/tlah18_public.pdf.

Leite, Bruno Scalia C. F.. 2022. "An introduction to mixed-integer linear programming: The knapsack problem." Towards Data Science, July 1, 2022. https://towardsdatascience.com/an-introduction-to-mixed-integer-linear-programming-the-knapsack-problem-1445452a9fe9.

Luszczyszyn, Dom. 2016. "Measuring Single Game Productivity: An Introduction To Game Score." Hockey-Graphs, July 13, 2016. https://hockey-graphs.com/2016/07/13/measuring-single-game-productivity-an-introduction-to-game-score/.

ODSC – Open Data Science. 2018. "Three Popular Clustering Methods and When to Use Each." Medium, September 21, 2018. https://medium.com/predict/three-popular-clustering-methods-and-when-to-use-each-4227c80ba2b6.

Sapra, Sunil K. 2010. "Robust Vs. Classical Principalcomponent Analysis in the Presence of Outliers." *Applied Economics Letters* 17, no. 6: 519–23. https://doi.org/10.1080/13504850802046989.

Schulte, Oliver, Zeyu Zhao, Mehrsan Javan and Philippe Desaulniers. 2017. "Apples-to-Apples: Clustering and Ranking NHL Players Using Location Information and Scoring Impact." Paper presented at: MIT Sloan Sports Analytics Conference, Hynes Convention Center, March 3 and 4, 2017. https://www.cs.sfu.ca/~oschulte/files/pubs/sloan-fix.pdf.

Steiger, James H. 2015. "Principal Component Analysis." StatPower, February 16, 2015. http://www.statpower.net/Content/312/R%20Stuff/PCA.html.

Stern, Alex. n.d. "Clustering NBA Player Types: A Tutorial on K-Means, Gaussian Mixture Models, Principal Component Analysis, and Graphical Networks." Github Pages (Blog), n.d. https://alexcstern.github.io/hoopDown.html.

Stimson, Ryan. 2017. “Identifying Player Styles with Clustering.” Hockey-Graphs, April 4, 2017. https://hockey-graphs.com/2017/04/04/identifying-player-types-with-clustering/.

T., Conor. 2015. “Using Cluster Analysis To Identify Player Position.” Hockey-Graphs, December 7, 2015. https://hockey-graphs.com/2015/12/07/identifying-player-position-using-cluster-analysis/.

Toledo Jr., Tiago. 2022. “PCA 102: Should you use PCA? How many components to use? How to interpret them?” Towards Data Science, January 4, 2022. https://towardsdatascience.com/pca-102-should-you-use-pca-how-many-components-to-use-how-to-interpret-them-da0c8e3b11f0#:~:text=Choosing%20the%20Principal%20Components,as%20possible%20of%20that%20threshold.