Identifying Player Archetypes in Women's Hockey

Carleen Markey[1] and Nayan Patel[2]

[1]*Undergraduate, Dept. of Physics and Astronomy and Dept. of Statistics, Purdue University, West Lafayette, IN*
[2]*Undergraduate, Dept. of Economics, The Ohio State University, Columbus, OH*

1. Introduction

Within men's hockey, player types, such as "offensive defenseman", "playmaker", and "goal scorer", are well established and prevalently known in and outside of the analytics community. However, in discussions surrounding women's hockey, there are no such firmly established archetypes that have not been simply borrowed from men's hockey and applied ill-fittingly. This is likely because large sample sizes of players and games did not exist in women's hockey until recently. Given this chance to work with the most in-depth set of data in women's hockey so far through the Big Data Cup, we attempted to find if player archetypes exist in women's hockey, and if so, what they are.

To identify these archetypes, we were inspired by astrophysics research one of the authors of this paper recently conducted, where she used characteristics of stars, the spatial coordinates, to identify groups of stars with clustering algorithms [3]. Because archetypes are also defined by a set of characteristics that a group of players is identified to have, we decided to use the same clustering methodology to identify groups of players that make up an archetype. After investigation, we found instances of similar approaches of finding player types in hockey [1,4,9,10], basketball [2], and soccer [7], which we used to affirm and guide our approach to this project.

2. Methods

2.1 Metric Calculation

Based on these previous clustering projects and our own domain knowledge, we identified eight measures of a player's contributions to generating offense: shots, primary shot assists, passes, successful entries, cross-slot passes, shots taken from the "home plate" area, takeaways, and puck recoveries. All of these metrics are counting statistics except for passing, which is a differential. We chose to measure passing in this way in order to potentially differentiate between true playmakers versus players whose passes may lead to lost possession. Because the Big Data Cup dataset lacked certain data, such as all of the players on the ice during an event, defensive metrics were much harder to extract and therefore excluded from our investigation.

We then created indexes for each of the statistics we identified to adjust each player's ability to be measured relative to their team, as done in Ryan Stimson's previous clustering work in 2017 [9]. This ensures that we can identify regardless if a player's team has dominated other teams or has performed poorly. An example formula for calculating the shot index for each player can be seen below in Eq. 1.

$$\text{Shot Index} = \text{Shots/Games Played} * (\text{Shots/Total shots taken by the player's team}) \qquad (1)$$

With all indexes calculated, we split the datasets into forwards and defenders to isolate archetypes within each position and then moved to the next aspect of our analysis to reduce the high number of metrics we have selected.

2.2 Principal Component Analysis

The more metrics used in clustering algorithms, the less likely the algorithm is to produce interpretable clusters. Principal Component Analysis (PCA) is a well-established method of addressing this problem [2,6]; therefore, we have chosen to use it in this project to reduce the number of variables. In short, this methodology creates a given number of new variables that are each equal to a unique linear combination of the existing variables. Applying this to our data, we found that we can preserve ~85% and ~99% of the information contained in the forward and defender datasets respectively by choosing to form 2 new PCA variables from the existing variables.

The resulting forward and defender datasets consist of player, player position, player team, and the two new PCA variables for each dataset. The first new PCA variable produced in the forward dataset prioritizes a player's passing, entry, and puck recovery abilities, which all factor into a player’s "playmaking" ability. Additionally, PCA forward variable 2 seems to measure shooting ability, with shooting, passing, entry, and danger shots all being important to that variable. For defenders, both new PCA variables emphasize puck recovery ability. However, the first PCA defender variable rewards player with a high passing volume and second PCA defender variable penalizes these players.

2.3 Clustering

With the forward and defender datasets from the principal component analysis, we then used three different clustering algorithms from scikit-learn [5], in order to identify the optimal one for this problem. Two of these algorithms, density-based spatial clustering of applications with noise and spectral clustering, did not identify any large, meaningful clusters at all. After further visual inspection of our data, we believe that this is due to the fact that both clustering algorithms are best used with data that has non-flat geometry, or a specific shape for each cluster. As shown in Fig. 3 and Fig. 4 later in this paper, our data after the principle component analysis does not appear to cluster into specifically shaped clusters, indicating it has a flat geometry.

Regardless, the third clustering algorithm tested, k-means, is suited to flat geometry and was able to identify clusters in the PCA data. There is also substantial evidence supporting the use of k-means in identifying player types [1,4,9,10]. To do this, we first set up Python code, found in the Appendix, to decide on the number of optimal clusters for the forward and defense datasets. We used two different methods to aid us in our decision, the elbow method as well as the silhouette coefficient.

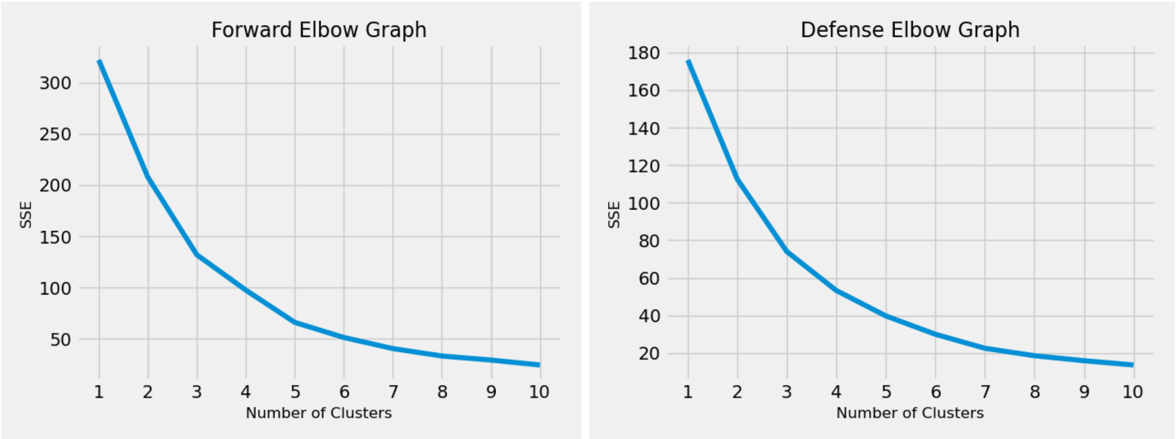

Figure 1: Elbow Graphs for the forward and defender datasets

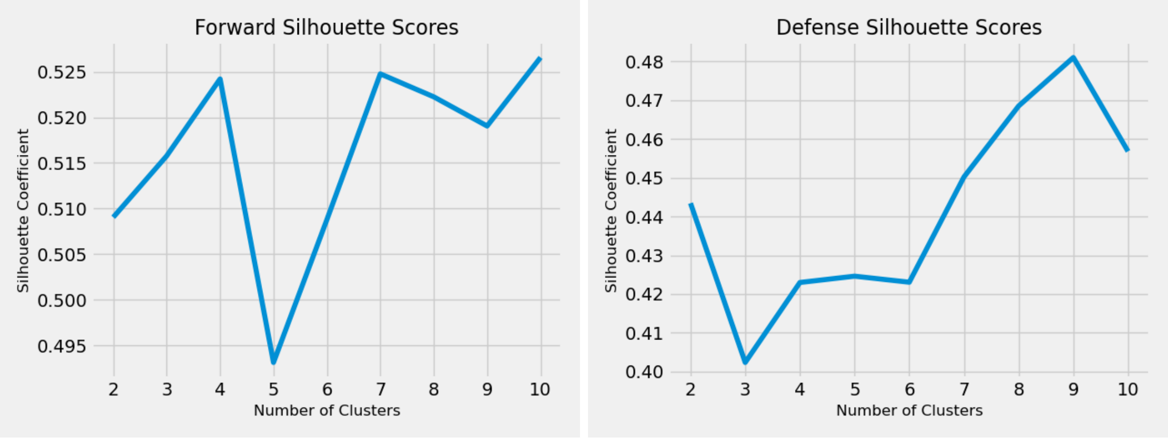

Figure 2: Silhouette Scores for the forward and defender datasets

Combining these two methods as well as incorporating our own knowledge on men's player archetypes, we decided that four clusters were optimal for forwards and three for defenders. Using these two numbers, we fitted our k-means model on both datasets using the "k-means++" initialization. This yielded a cluster designation for each player in our dataset.

3. Discussion

After fitting the k-means model to our forward PCA data, we summarized each cluster by their indexes to assign archetypes and descriptions:

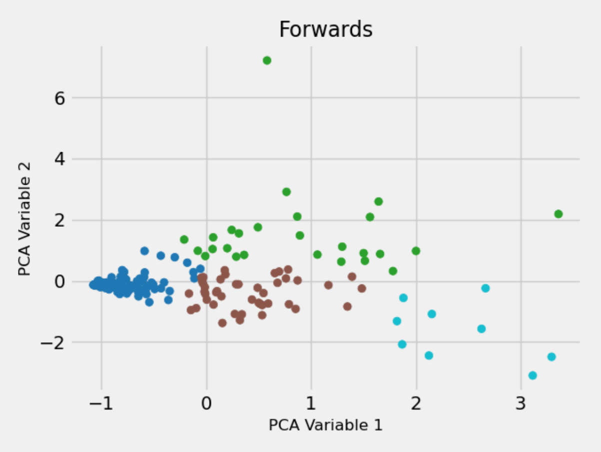

Figure 3: Output of Forward PCA k-means model, clustered by color

Forwards									
	Shot Index	PSA Index	Passing Index	Entry Index	Danger Passing Index	Danger Shot Index	Takeaway Index	Puck Recovery Index	Count
Dependent	0.063	0.027	0.208	0.083	0.042	0.05	0.052	0.168	82
Balanced	0.243	0.226	1.098	0.292	0.222	0.188	0.158	0.687	44
Shooter	0.643	0.285	1.25	0.794	0.389	0.518	0.145	0.821	26
Playmaker	0.413	0.626	2.75	0.597	0.467	0.278	0.312	1.482	9

Dependent (Dark Blue - Figure 5)
Sarah Nurse, Meghan Duggan, Tori Sullivan, Amy Curlew

Dependent players mostly fill out rosters, excelling in no particular metric. They tend to rely on others to generate offense while maneuvering themselves into open positions to receive/give passes to continue their respective offensive systems. While they tend not to individually stand out offensively, they play an important "cog in the machine" role within the tactical setup, and without them, the structure can fall apart.

Balanced (Brown - Figure 5)
Melodie Daoust, Amanda Kessel, Jillian Dempsey, Natalie Marcuzzi

Balanced players are very similar to Dependent players in that they don't excel in any particular one metric. However, they do exceed Dependent players in generating and creating offense. These players are the ideal depth players to fill out the bottom six and the holes in the lineup where Playmakers or Shooters otherwise cannot be added. They also help with the transition and defensive aspects of the game generating a higher number of takeaways than Dependents or Shooters.

Shooter (Green - Figure 5)
Kendall Coyne Schofield, Élizabeth Giguère, Kelly Babstock, Mikyla Grant-Mentis

Shooters have the highest shooting and danger shooting indexes, obviously, with strong (but not elite) play driving and contribution numbers. These players tend to find the open areas on the wings and in the slot by carrying the puck into the zone and coupled with elite shooting talent, find the back of the net on a regular basis. As a matter of clarification, Valeria Pavlova actually was given her own cluster by the model due to her shooting numbers being way higher than any other in the Stathletes data. We decided to merge her cluster with the Shooter cluster, as it is nonsensical for an archetype to be made up of one player, and her metrics were most similar to this group

Playmaker (Light Blue - Figure 5)
Marie-Philip Poulin, Brianna Decker, Shiann Darkangelo, Madison Packer

For our final archetype, we have what has been, analytically, determined as the most valuable offensively in modern hockey tactics. These players have the highest play driving and puck recovery numbers. They are the quarterbacks of their teams' offensive systems and have a knack for finding open players in dangerous areas to create opportunities. Our model only identified nine of these players in Stathlete's datasets, hence the inherent value in having as many of these players in your roster.

To compare the clusters we found to Ryan Stimson's men's hockey archetype project [9] to see what the differences were between men's and women's hockey. For forwards, we found that the archetypes are largely the same between genders, showing that offensively, players fill similar roles to score.

For defenders, we found three clusters to be optimal, with interesting results:

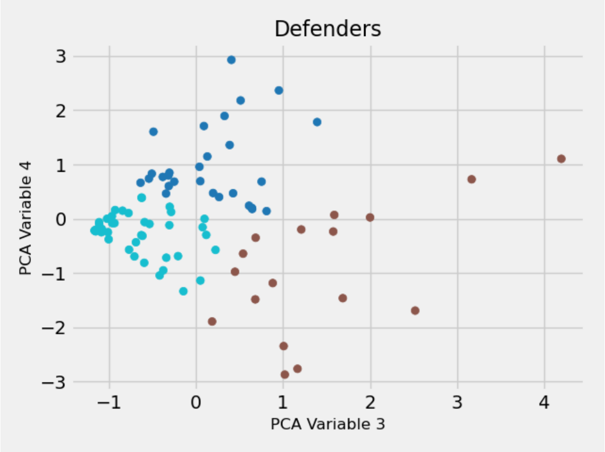

Figure 4: Output of Defender PCA k-means model, clustered by color

Defenders									
	Shot Index	PSA Index	Passing Index	Entry Index	Danger Passing Index	Danger Shot Index	Takeaway Index	Puck Recovery Index	Count
Defensive	0.076	0.021	0.819	0.018	0.019	0.012	0.08	0.611	43
Disruptor	0.188	0.073	1.913	0.085	0.039	0.016	0.203	2.052	28
Two-Way	0.504	0.222	4.453	0.126	0.107	0.021	0.26	3.128	17

Defensive (Light Blue - Figure 6)

Jamie Bourbonnais, Minnamari Tuominen, Sidney Morin, Lisa Chesson

This archetype makes up a majority of the defenders in the dataset. These players, understandably, don't excel in any of the offensive metrics used in the clustering. Takeaways and puck recoveries are the closest thing we had to tracking defensive play, and even then, they are more transitional stats in the authors' eyes. We know that players in this archetype tend to play deeper in their own zone, and don't tend to be active on the rush, preferring to support breakouts.

Disruptor (Dark Blue - Figure 6)

Paige Capistran, Tori Howran, Whitney Dove, Saroya Tinker

The disruptor archetype is one that we think is really unique to women's hockey. These players are similar to defensive defenders, except for having much higher takeaway, puck recovery, and passing numbers. Because of the slower rushes and increased passing in the offensive zone compared to men's hockey, it creates a need for defenders to play a more active role in the defensive zone clogging up passing lanes and pressuring the puck. This leads to those higher puck recovery and takeaway numbers without increasing the offensive metrics, since they still play a more defensive role on the breakout and in the offensive zone.

Two-Way (Brown - Figure 6)

Lee Stecklein, Emily Pfalzer, Taylor Turnquist, Lindsay Eastwood

The two-way defender archetype is on the other end of the spectrum with these players activating more on the rush and playing a bigger role in their team's offensive zone setup. The biggest jump seen is in the passing and shooting indexes compared to their positional peers, which shows their tendency to be more involved with creating offense.

When comparing our findings to Ryan Stimson's men's hockey archetype project [9] for defenders, we do actually see that women's hockey defenders play different roles compared to the men's game. We postulate that this is due to differences in the way the game is played between genders, research that one of the authors of this paper recently conducted [8]. Women's hockey tends to see more setup play in the offensive zone and slower, more thoughtful puck movement. Men's hockey has more goals scored from off the rush and quick plays. This difference would definitely change the way that defenders play the game, something we see in our clusters.

4. Conclusion

The biggest takeaway we can take from this project for women's hockey coaches, scouts, and general managers is that this is a very helpful first step in data-driven roster construction and lineup optimization. In Ryan Stimson's article, he uses expected goals data to show what player archetypes to target for a team, as well as what archetypes mesh the best with others when setting your lineup [9]. This kind of thought process can help a team target draft picks and free agents (or recruit) more efficiently, as well as pair together complementary players within forward lineups or defender pairings to maximize their skillsets.

This can also be helpful for scouting purposes, when game-prepping for opponents. Coaches can see the type of players to target on the other team to defend and play closer. Instructing players to pay closer attention to shooters and playmakers in the defensive zone can help cut down on high danger passes and shots.

In conclusion, we have identified data-driven archetypes in women's hockey, some of which are new and some of which have been previously established in men's hockey. Particularly, in comparing men's and women's hockey, forward archetypes are very similar across the two types of hockey, but critical nuances cause defender roles to be different.

Acknowledgments: We would like to thank Ryan Stimson for his clarifications on his 2017 article, Matt Schulz and Dr. John Draper for aiding in editing, and Stathletes for providing this in-depth dataset for this project.

References

[1] A. Novet, Goal Scorer Cluster Analysis, https://hockey-graphs.com/2018/01/04/goal-scorer-cluster-analysis/ (2018)

[2] A. Stern, Clustering NBA Player Types: A Tutorial on K-Means, Gaussian Mixture Models, Principal Component Analysis, and Graphical Networks, https://alexcstern.github.io/hoopDown.html (Unknown)

[3] C. Markey, D. Guszejnov, and S.S.R. Offner, Origins of Mass Segregation in Stellar Clusters within the STARFORGE Simulations, *Research Notes of the American Astronomical Society,* **4,** 163 (2020)

[4] C. Tompkins, Using Cluster Analysis To Identify Player Position, https://hockey-graphs.com/2015/12/07/identifying-player-position-using-cluster-analysis/ (2015)

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, **12**, 2825-2830, (2011)

[6] J. Lever, M. Krzywinski, N. Altman, Principal component analysis, *Nature Methods*, **14**, 641–642 (2017)

[7] J. Muller, The Seven Styles of Soccer, https://spacespacespaceletter.com/the-seven-styles-of-soccer/ (2021)

[8] N. Patel, Using Shot Tracking & Transitional Play Data to Quantify the Systemic Differences Between Men's and Women's Hockey, https://www.hockeyuanalytics.com/blog (2020)

[9] R. Stimson, Identifying Playing Styles with Clustering https://hockey-graphs.com/2017/04/04/identifying-player-types-with-clustering/ (2017)

[10] T. C. Chan, J. A. Cho, and D. C. Novati, Quantifying the Contribution of NHL Player Types to Team Performance, *INFORMS Journal on Applied Analytics,* **42**, issue 2, 131-145 (2012)

Appendix - Code

Fully structured code and generated data can be found at

https://github.com/cmarkey/Player-Clustering