**NHL Draft Analysis**

I have decided to do a clustering analysis on the careers of all players drafted in the NHL Draft. As hockey is my favorite sport and have an interest in following the NHL Draft, I feel that doing an analysis on the NHL Draft is fitting. My goal of this analysis is to see which positional players and goaltenders are clustered together. I want to see if there is a general trend between the careers of certain players. For example, there can be clusters that are only exclusive to players that are in the Hall Of Fame. It would also be interesting to see how important a statistic is for a certain cluster, for example, does the number of games played make a difference for a cluster. **(All plots and graphs are at end of paper)** The dataset contains each player drafted and their career statistics. The dataset goes from the first NHL Draft up until the 2022 NHL Draft, so all data from the 2022-2023 NHL season will not be included. The unique part of the analysis is that there will be two analyses, as one analysis is a clustering of the positional players, which are centers, right wingers, left wingers, and defensemen. There are also players that have secondary positions in the position column, for example, a player that has the position of “C/LW” means that they played both center and left wing during their playing career. Some variables that I plan on using are goals, assists, and points, which are important for positional players. I have also made variables for whether the player is in the Hockey Hall Of Fame, whether they played 100+, 500+, and 1,000+ games, and if they hit 500+ points and over 1,000+ points. For goalies, some variables that I plan on using are goals against average and save percentage. I am also using variables that I created including if the player is in the Hockey Hall Of Fame, whether they hit 100+ wins and whether they hit 350+ wins. For both positional players and goalies, I have created a variable to show whether that player was a top-15 pick in their draft. My goal of this analysis is to see which positional players and goaltenders are clustered together. I want to see if there is a general trend between the careers of certain players. For example, there can be clusters that are only exclusive to players that are in the Hall Of Fame.

**Data Cleanup**

In order to make two datasets that are optimal, I had to decide on which variables to use for both the position player and goalie clusters. I made two datasets off of the original dataset, with the position player dataset having the filter, “position != “G””, which removes all rows where the position is G for goalies. For the goalies dataset, I used the same filter, but instead of “!=”, I used “==” which contains all rows that have the position “G.” There were null values in the goaltender dataset, so I used NHLGoalies\_999 <- replace(NHLGoalies, is.na(NHLGoalies), 0) to replace all NA values with 0. After I created the dataset and removed the NA values, I created a subset of each variable that I thought would be valuable for the analysis. For the players cluster, I used goals, assists, points, points per game, Hall Of Famer, 100+ games played, 500+ games played, 1,000+ games played, 500+ points, 1,000+ points, and top 15 pick. For goalies, I used Hall Of Famer, save percentage, goalie wins, 100+ wins, 350+ wins, goalie losses, goals against average, games played, and top 15 pick.

**Player Clusters**

When looking at the elbow chart for positional players, the bend starts at 4, meaning that 4 clusters is optimal for the analysis. To get the optimal number of clusters, I used fviz\_nbclust(NHLCluster, kmeans, method = "wss").

To visualize the cluster plot, I first used clusterP <- kmeans(NHLCluster, center = 4, nstart = 50) to generate the clusters and get the cluster means. To visualize the clusters, I used fviz\_cluster(clusterP, data = NHLCluster). On the cluster plot, clusters 1, 2 and 4 are all close to each other, with many of the points being in the same general area. Cluster 3 has many points in the general area as the other clusters, but there are many points that are scattered throughout the area of the plot. When looking at cluster 1, there is one point that is an outlier. This point represents Bill McCreary, who was a NHL Draft pick, but was inducted for his work as a referee. On clusters 2 and 4, the outlier points are players that are Hall of Famers, there are 7 points in cluster 2 and 3 in cluster 4, 2 of the points in cluster 4 are on top of each other, as the player was drafted twice. When looking at cluster 3, there are 62 values that are distanced away from many of the points in the cluster, and the points that are away are Hall Of Famers. When looking at the mean values of each cluster, cluster 2 and 3 stand out from the other 2 clusters. Both cluster 2 and 3 have the highest means of the 4 clusters, with cluster 3 having a sizeable mean difference over cluster 2. The difference in means can be attributed to both the number of players in each cluster as well as the type of players in each cluster. When looking at cluster 1, there are 8,891 players that are put in the cluster. The players that are in the cluster are either: young players that were recently drafted and have not played a game in the NHL yet, were drafted and never played in the NHL, are young players that have played in the NHL, or have had short careers. When looking at the cluster means, there is an average of 17.16 games played, 1.58 goals, 2.74 assists, and 4.33 points. When looking at who has the highest values, Andrew Peters and Greg Smyth both played 229 games, which is a short career. The leader in goals, assists, and points, is Kirill Kaprizov, who started his NHL career in the 2020-2021 NHL season. Granted, he, along with many players in the cluster, are young and will get more games played, goals, and assists in their career. There is one Hall Of Famer who played 12 games in the NHL, and was inducted into the Hall Of Fame as a referee. There were only 621 players in the cluster that have played at least 100 NHL games, which is 6.21% of the whole cluster. The cluster also has the most top 15 picks, but they make up 3.22% of the entire cluster. When looking at cluster 2, there are 781 players in the cluster. The players in the cluster are either: stars in the NHL currently, Hall Of Famers that had short careers, or players not known for their offensive skills. The mean for games played is 809.03, and points is 388.8. The games played leader is Luke Richardson with 1417 games played, who was known for his defensive and physical play style. Pavel Bure has the most goals(437) and points(779) in the cluster, and was known for his speed and offensive skillset. He had a short career, but is in the Hall Of Fame. Connor McDavid has the highest points per game with 1.43, and is currently the best player in the NHL, and will go down as one of the greatest NHL players in history. There are 7 Hall Of Famers in the cluster, which makes up 0.9% of the cluster. The cluster has the most players that have played 500+ games at 777 and is 99.49% of the cluster. There are 193 players that have hit 500 points, which makes up 24.71% of the cluster. 31.63% of the cluster is a top 15 pick. When looking at cluster 3, the means are the highest for each variable. The average number of games played is 1,170.9 and the average number of points is 946.58. The players that have the most goals, assists, points, games played, and points per game and are some of the all-time greats in NHL history. 27.56% of players in the cluster are in the Hall Of Fame, all players have at least 500 games played and 500 points, and there are 80.44% of players that have played at least 1,000 games and 33.33% that have at least 1,000 points. 53.78% of the players in the cluster are top 15 picks. When looking at cluster 4, there are 1,136 players in the cluster, and contains a mix of current NHL superstars, current NHL players that are in the middle-six and top-4 defensemen, and middle-six forwards that have had short careers. Ryan Reaves has the most number of games played in the cluster with 755, and his known for his physical game. Auston Matthews has the most goals(259) and points(457) in the cluster. He was drafted in the 2016 NHL Draft and is one of the top goal scorers in the NHL. He is the first American to hit 60 goals in a single season. Every player in the cluster has played at least 100 games and 27.2% of players have played at least 500 games. 17.87% of the players in the cluster are top 15 picks.

**Cluster Comparisons**

For the analysis, there was some work that I had to do beforehand. I used NHLDraft\_1.labels = NHLDraft\_1$player to have the labels be the row names. I then used rownames(NHLCluster) <- paste(NHLDraft\_1$player, 1:dim(NHLDraft\_1)[1], sep = "\_") to get each player’s name with a number next to it as the row labels, an example being Alexander Sammarco\_123. I then exported each cluster information with this code: write.csv(NHLCluster[clusterP$cluster==1,], "C:/Users/asamm/Documents/DSS 665/Cluster1Players.csv", row.names = TRUE, col.names=TRUE), After that, I did a column split by the \_ and the number after the player’s name and removed the underscore, and had the number be the player identifier. When looking at some comparisons between players in clusters, I found that the games played and points swayed which cluster a player was in. For example, Ryan Reaves(cluster 2) and Rob Ray(cluster 4) are comparable. Rob Ray has 900 games, while Ryan Reaves has 755 games played, even though Ryan Reaves has more points. While the difference in points between the two is small, the games played is the difference. Another example is Glen Wesley (cluster 3), Luke Richardson (cluster 4) and Darryl Sydor (cluster 4). While Luke Richardson only has 30 less games than Glen Wesley, Wesley has over 300 more points. For the comparison of Sydor with Wesley, both have near the same number of points but Wesley has close to 300 more games played. One more example is Pavel Bure (cluster 2) and Peter Forsberg (cluster 3), as both have around the same number of games, but Forsberg has over 100 more points than Bure.

**Goalie Clusters**

When looking at the elbow chart for positional players, the bend starts at 3, meaning that 3 clusters is optimal for the analysis. To get the optimal number of clusters, I used fviz\_nbclust(NHLGoalies\_123, kmeans, method = "wss").

To visualize the cluster plot, I first used clusterG <- kmeans(NHLGoalies\_123, center = 4, nstart = 50) to generate the clusters and get the cluster means. To visualize the clusters, I used fviz\_cluster(clusterG, data = NHLGoalies\_123). On the cluster plot, there are a couple of points that stand out for cluster 1. The point at the bottom of cluster 1, 421, is Martin Houle, who played 1 NHL game, had a save percentage of 0.667, and had a goals against average of 27.27. Point 1123 is Ken Holland, who is in the Hall Of Fame for his work as an NHL executive, was drafted and played a couple of games in the NHL. On cluster 2, there are 5 points that are separated from the rest of the points, and they are 5/6 Hall Of Famers in the cluster. The 6th Hall Of Famer in the cluster is grouped with the few players that are between the 5 Hall Of Famers and the rest of the goalies in the cluster. When looking at cluster 1, there are 1,033 goalies that are in the cluster. Many of the goalies in the cluster have either never played a game, are recently drafted and haven’t played a game, young goalies that have not played many games in the NHL, and goalies that were brought up from the minor leagues to play a couple games. There is one Hall Of Famer in the cluster, but was inducted for his work as an executive in the NHL. Jack Campbell has the most wins (71), and is currently a starting goalie in the NHL. Tommy Soderstrom has the most games played, he was a backup goalie in the 1990’s. The goalies that have a save percentage of 1 have either played only 1 game and didn’t let in any goals or replaced the starting goalie in a game due to poor performance or injury. The 745 goalies that have 0 goals against average have never played a game in the NHL. The cluster has the lowest goals against average, but this number is lower than the other clusters for the fact that there are many goalies that have not played a game in the NHL. Similar to cluster 3 for the positional players, cluster 2 has some of the all-time great goalies. There are 60 goalies in the cluster, and 10% of them are Hall Of Famers. Every goalie in the cluster has won at least 100 games, and 28.33% of the goalies have won at least 350 games. When looking at the cluster means, the cluster has the highest average save percentage, most wins, most 100-win and 350-win goalies, the second lowest goals against average (lowest if not counting cluster 1), and the most games played. When looking at the largest values, there are 2 goalies that are the leaders for multiple categories. Dominik Hasek has the highest save percentage (0.922) and lowest goals against average (2.2, lower is better for this statistic), and Martin Brodeur has the most wins (691) and most games played (1,266). Cluster 3 has 124 goalies. Many of the goalies in the cluster are career backup goalies, actively playing, or had short careers as starting goalies. The cluster has the 2nd highest save percentage of 0.897, the second most wins with 126, but has the lowest goals against average with 3.04. There are 2 Hall Of Famers in the cluster, which makes up 1.61% of the goalies. One of the goalies, Jim Rutherford, played 457 games, but was inducted into the Hall Of Fame for his work as an executive. Ken Dryden is the other Hall Of Famer and had a record of 258 wins and 57 losses. He played a total of 8 seasons and became a professor and a member of congress in Canada. Dryden also has the highest save percentage of all goalies in the cluster. Frederik Andersen has the most wins in the cluster with 261 and is actively playing in the NHL. Jordan Binnington has the least number of losses (46) and is currently playing. Steve Mason and Glen Hanlon both have the most games played with 476, both goalies were starters at some point in their careers, then became backups as they got older. **Cluster Comparisons**

To set the player comparisons up, I also used the methodology that was used for the player analysis, the difference being the dataset names. One comparison for the goalies is between Steve Mason (cluster 3), Glen Hanlon (cluster 3), and Corey Crawford (cluster 2). While the 3 have close to the same number of games played, Crawford has more wins, a better save percentage, a better goals against average, and less losses. Another example is Frederik Andersen (cluster 3) and Jeff Hackett (cluster 2). Even though Andersen has better stats overall, Hackett played 55 more games, which puts Hackett in cluster 2 while Andersen is put in cluster 3.

**Final Comments**

After running both cluster models, I have a better understanding of the types of players that would be clumped together. While it makes sense for the best players to be in the same cluster, I wanted to see where current players would be placed. When looking at the position players, there was always there was great variation between each the elements in a cluster. In the first cluster, it was players that have played 250 games or less, as well as all the players that were drafted that never played a game in the NHL. In the 2nd cluster, there were players that played had less points than the top point producers in the 1st cluster, but they played a lot more games. Cluster 3 is the cluster with the best players that have been drafted, and the cluster means show. The 4th cluster is similar to the 2nd cluster, as there are current NHL players and other players that are less known for their offensive abilities. For the goalies, cluster 1 contains goalies that have played few games in the NHL or have never played in the NHL. Cluster 2 contains the best goalies in NHL history along with goalies that have had long careers. Cluster 3 contains goalies that were career backups or had short careers as starting goalies. A question that I have is seeing whether the clusters would be different if many of the players that never played a game were excluded. Also, this analysis would be interesting to go back to every couple of years to see the difference in clusters between players.

**Graphs and Descriptive Statistics on Datasets**

**Clusters**

**Player Clusters**

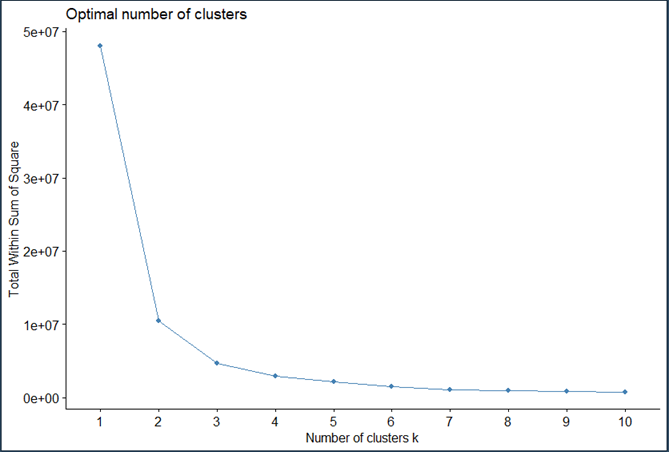
A graph with a line

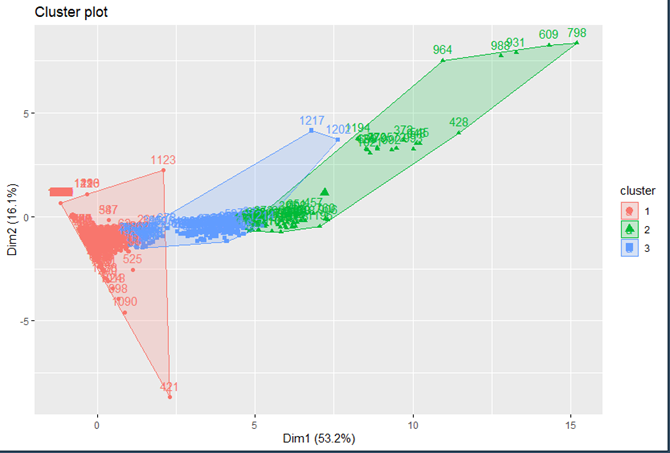
Description automatically generated

A graph showing different colored lines

Description automatically generated

**Goalie Clusters**

****

****

**Player Statistics (Data below graphs, 1 = Yes, 0 = No)**

**Games Played**

A graph of a game

Description automatically generated

Mean: 137.15

**Goals**

A graph with a bar and text

Description automatically generated

Mean: 24.18

**Assists**

A graph of a number of individuals

Description automatically generated

Mean: 40.23

**Points**

A graph with a bar and a number of points

Description automatically generated

Mean: 64.41

**Points Per Game**

A graph of points per game

Description automatically generated

Mean: 0.13

**Hall of Fame**

A blue square with white text

Description automatically generated

0 1

10960 73

**100+ Games**

A blue squares with white text

Description automatically generated

0 1

8270 2763

**500+ Games**

A blue squares with white text

Description automatically generated

0 1

9722 1311

**1000+ Games**

A blue square with white text

Description automatically generated

0 1

10734 299

**500+ Points**

A blue squares with numbers

Description automatically generated

0 1

10615 418

**1000+ Points**

A blue square with white text

Description automatically generated

0 1

10958 75

**Top 15 Pick**

A blue square with white text

Description automatically generated

0 1

10176 857

**Goalie Statistics (Data below graphs, 1 = Yes, 0 = No)**

**Goalie Hall Of Fame**

A green square with black text

Description automatically generated

0 1

1208 9

**Save Percentage**

A graph with green bars

Description automatically generated

Mean: 0.36

**Goalie Wins**

A graph with a green bar

Description automatically generated

Mean: 31.27

**100+ Wins**

A green squares with white text

Description automatically generated

0 1

1079 138

**350+ Wins**

A green square with numbers

Description automatically generated

0 1

1200 17

**Goalie Losses**

A graph with green bars

Description automatically generated

Mean: 27.75

**Goals Against Average**

A graph with a green bar graph

Description automatically generated

Mean: 1.36

**Goalie Games Played**

A graph of a game

Description automatically generated

Mean: 73

**Goalie Top 15 Pick**

A green squares with numbers

Description automatically generated

0 1

1179 38

**R Reflection**

Throughout the 8 weeks of the course, I have learned a lot about R, and have found it to be very interesting. As someone who uses Python at work, I found the learning curve to be easy. I felt that learning for-loops was difficult, as the brackets can be difficult to remember, but I find that the + symbol is helpful when doing the loops. Unlike Python, R gives more direction and makes it easier for guidance. I enjoyed learning about the anova and multiple linear regression models, as my goal is to be a data scientist once I am done with school, and regression models are a large part of the work that data scientists do. When running the model, it’s important to see which variables are significant, and when seeing the p-values for the variables in R, I am able to see a star next to each of the p-values. I also liked learning about the visuals in R. Compared to Python, I find that the visualizations in R are cleaner than they are in Python. I also like the ability to zoom in to the point where you can have the visualization be the whole screen. I wish that I was able to learn how to use more machine learning and predictive modeling in R. At work, I have built Word2Vec models using Python, and I have also built TensorFlow models on my own time in Python. I would like to see the differences in how the coding is done with machine learning models in R. Also, I would like to see how some of the models look visually in R. It would be interesting to do a data science project workflow, as I can get the data, clean the data, initially run the model, make model changes, then present the model and why variables are important to the model’s success, as it can be something that I can learn for my future. Overall, I feel that this course has really helped my coding skills, not only by learning R, but I can translate some of the things I learned in R and use in Python.