CSCI 4022 Final Project Write-Up

Jacob Reynolds

Every professional sport around the world has a small set of athletes who are simply better than the rest. They are continuously putting up ridiculous statistics, break numerous records, and making historic moments against the best competition in the world. Their talents and abilities are unquestionably the greatest in the world today. But how do they compare to other greats in their respected sport from previous generations? Would they still dominate their sport if they played in a different era? The same question could be asked from the other perspective. Would a Hall of Fame level athlete from the 1980s be able to keep up and display the same brilliance they exemplified in today's game? Every professional sport enthusiast has asked the same question, "Who is the G.O.A.T?" (Greatest of All Time). In the world of professional hockey, Wayne Gretzky and Mario Lemieux are two names who are always the central point of this conversation. The numbers that they put up back in the late 1900s are simply remarkable, and no player since has come close to matching their points per game statistics from their primes. But the game has changed since they both played in the NHL. Goaltending, for one thing, has seen a massive improvement due to the improvements in goalie equipment and technical play style. This has caused the average goals per game benchmark to drop by approximately 3- 4 goals since the Gretzky or Lemieux era. This is just one example for why it is so hard to compare 2 athletes who played in different eras. If you look at the NHL today, players like Connor McDavid, Sidney Crosby, and Nathan MacKinnon consistently make fools of their opponents and show exemptional abilities that the NHL has never seen before. But they never seem to be given as much attention as some of the older greats of the league in the G.O.A.T discussion simple because their statistics, even whilst being the best by today's standards, do not compare to the history of the league. Are statistical records of the past a biproduct of the goal-frenzy game style played back in the day? Or were the Hall of Fame legends of the past simply ahead of their time and an anomaly that can't be seen in today's game? So, to try and answer this age-old question of, "Who's the G.O.A.T?", I want to try and use some advanced data science techniques to attempt to better compare players from different eras.

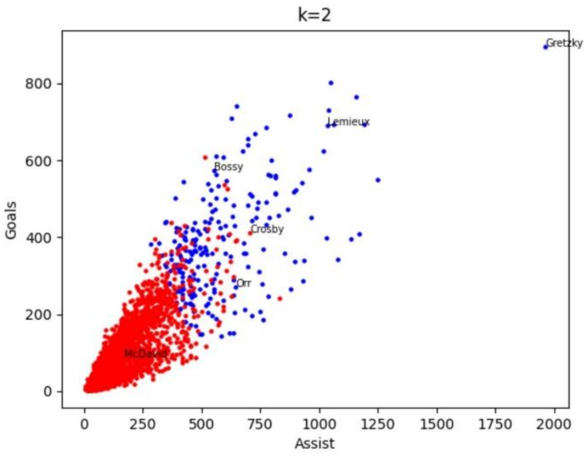
The data I used was an extensive list of a multitude of statistics of every player from every single season of the NHL from 1940 to 2018. This means if John Smith played 10 seasons in the NHL from 2003 to 2013, then this data set would have 10 different entries labeled "John Smith", with his individual statistics for each of his 10 seasons. Multiply this by the number of players who have ever played in the NHL from 1940 to 2018 and you get a data set of 37,826 individual entries. This excel sheet was provided from a very large Kaggle submission of a million other NHL History statistics. Although having each player's individual season statistics was nice, I wanted each player's career statistics. This required me to do an extensive amount of data scrubbing with python as I created a new data set of each player's career statistics. I also didn't need every statistic provided from the original data, so I scrubbed until I was left with just the following career statistics for each player.

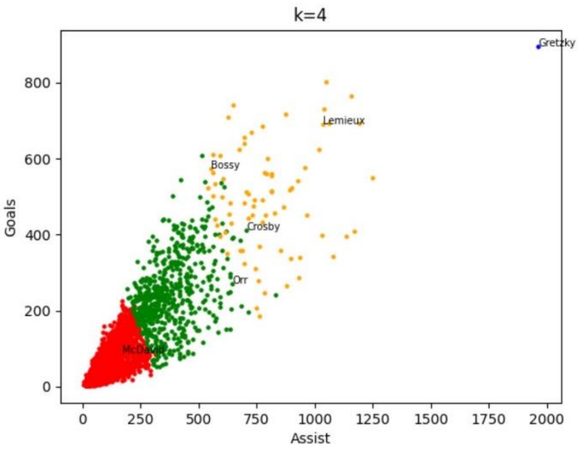
- Name
- Position
- Games Played
- Goals
- Assists
- Points
- Plus/minus
- Shots
- Shot Percentage
- Goals per Game
- Assist per Game
- Points per Game
- Plus/minus per Game
- Shots per Game

These statistics were chosen because they accurately provide statistical reasoning towards a player's value and effectiveness on the ice. Statistics including blocked shots, hits, Faceoff percentage, and shorthanded/powerplay goals were left out because they either do not correlate to how good a player's skills are or are too niche of a statistic to be compared across players with different positions. Blocked shots at times can be random. Hits are usually accumulated by role-oriented grind players whose job is to be a physical force on the ice rather than a goal scorer. Faceoff percentage can only be calculated by players who play the center position. And shorthanded/powerplay goals are too niche of a statistic in my opinion to be considered valuable when comparing players across different positions, teams, and eras. Statistics like time on ice and offensive/defensive time on ice were left out because those statistics were not calculated for players that played back in the 1970s and prior. I feel that those stats would've been a nice addition to have been calculated but it wouldn't have been fair for older players as I would've had to include a 0 placeholder and newer players would've had an advantage. My last step of scrubbing including removing players who didn't play a minimum of 200 game. This was necessary because playing 200 games gives a player enough time to have reliable enough data that displays an accurate representation of that player. Thankfully, all this scrubbing reduced my data set from 37,000 entries to just about 2,600 entries. The full Kaggle data set can be found here: https://www.kaggle.com/open-source-sports/professional-hockey-database?select=Master.csv

Anyone who follows/studies the NHL knows that Wayne Gretzky is the greatest of all time with respect to statistical records. By the time he retired in 1999 he held 61 individual NHL records, all of which he still holds today. This includes records like most NHL career goals (**894**), most NHL career assist (**1,963**), most NHL career points (**2,857**), most NHL points in one season (**215**), most 100-point seasons (**15**), most 50-goal seasons (**9**), and so on. While these records are impressive, there are a lot of players whose achievements and skills go under looked due to multiple condition. Here are a few examples of great NHL players who at face value don't compare to Gretzky. Mario Lemieux's career statistics rank him 7th in career NHL points. But on average, Lemieux played 500 less games than everyone else in the top 10 due to injuries. Bobby Orr similarly is ranked 112th in career points but only played 657 games due to injury. That's on average 800 less games than players in the top 10. Some current NHL players like Sidney Crosby and Connor McDavid are the best in the league right now and haven't finished their careers yet. It's not fair to compare their stats to that of someone who has already played an entire NHL career's worth of games. Given these examples and conditions, you could make an argument for anyone of these players to contend with Gretzky for the G.O.A.T title. Many sports analysist have already done so, and the debate is never ending as there are so many variables to look and so many candidates to consider.

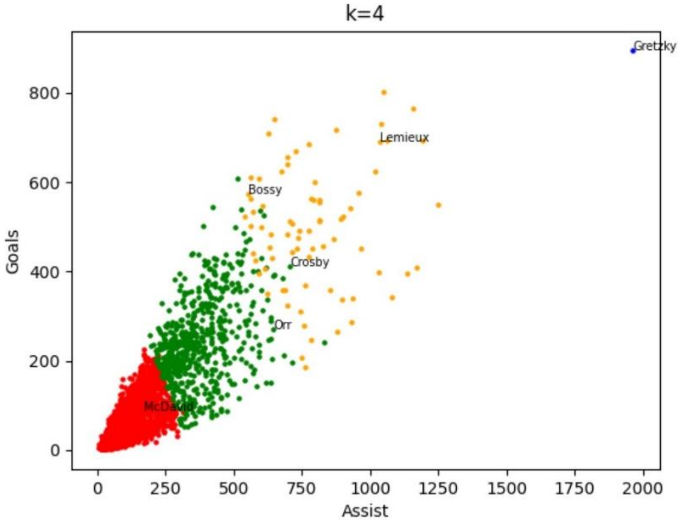
I have taken my scrubbed data set and run the k-means clustering algorithm we learned in class on it to see if we can separate the best from the rest. I have decided on a k value of 4 has worked best to create clusters small enough to allow the number of players in the tier 1 cluster to be a suitable size. If we have 100 players listed as candidates for the G.O.A.T title, then either our statistics being measured are too generic or we don't have enough clusters. Either way, our discussion would be pointless trying to compare 100 players who advanced statistics are all relatively similar. Below is a comparison example of the clustering difference between k=2 and k=4.



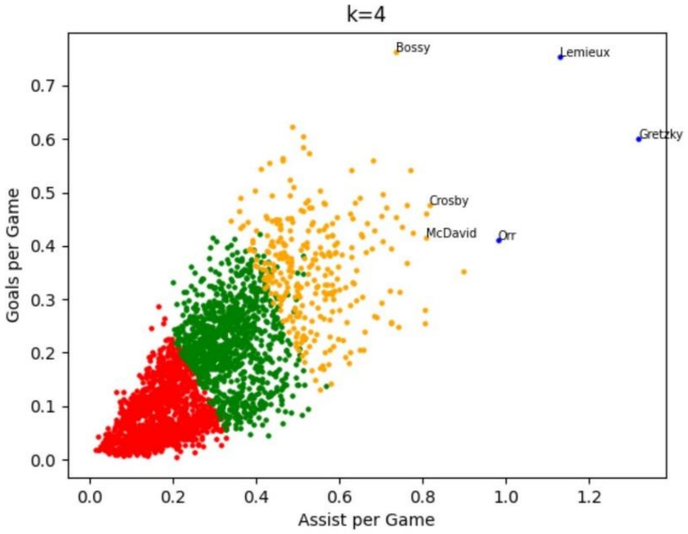


Because our goal prioritizes comparing the best of the NHL to the rest, the initialization of the first centroid has been assigned to the best player with respect to the statistics being measured. This is to ensure a cluster includes just the top players, the question is just which other players join him. The remaining cluster centroids are generated randomly. The rest of the algorithm processes through each player and calculates its Euclidean distance from each of the cluster's centroids using the designated statistics of the player and centroid. The player is then added to the cluster whom it is closest to and the centroid of that cluster is then recalculated.

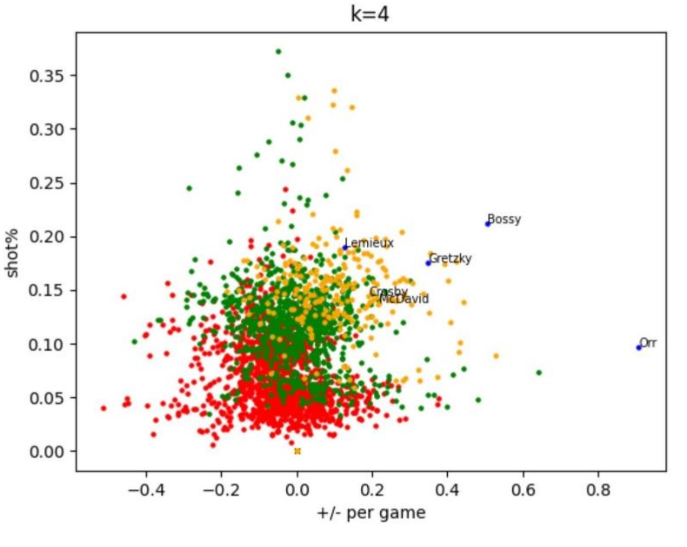
As you can see below, measuring career goals vs. career assists demonstrates exactly what we explained before. Wayne Gretzky's career numbers are just in a league of its own as he is the only member of the tier 1 cluster. This checks out because Gretzky has a 93 more goals than 2nd place (Gordie Howe) on the all-time NHL goals leaderboard and 714 more assists than 2nd place (Ron Francis) on the all-time NHL assists leaderboard.



Using more advanced statistics like goals per game and assists per game, displayed below, gives us a much different graph structure. Here we can see that Wayne Gretzky is joined by Mario Lemieux and Bobby Orr in the tier 1 cluster. This checks out because Lemieux averaged 0.75 goals per game (0.15 more than Gretzky) and 1.13 assists per game (0.19 less than Gretzky) and Orr averaged 0.41 goals per game (0.19 less than Gretzky) and 0.98 assists per game (0.34 less than Gretzky). Looking at the graph however, I would include Mike Bossy as an honorable mention for this cluster as he did average 0.76 goals per game (0.16 more than Gretzky) and 0.74 assist per game (0.56 less than Gretzky).



Lastly, I attempted to add some additional advanced statistics in our calculations to see if we can find any tier 1 level players with stats other than just goals and assist. In the graph below I used goals per game, assists per game, plus/minus per game, and shot percentage to calculate Euclidean distances in the 4th dimension. Because graphing in this dimension is visually impossible, I have just displayed the shot percentage vs. plus/minus per game graph but with the correct clusters colored appropriately. Here the algorithm gave us the same tier 1 cluster from the previous test, but with the addition of Mike Bossy. This feels like a more complete list than before because in the previous graph you could see Mike Bossy's data point clearly separated from the rest but was just shy of joining the tier 1 cluster.



In conclusion, k-means clustering did an effective job of separating the players of the history of the NHL into respectful tiers based on the measured statistics. With this algorithm we were able to accurately derive a new list of players that should be considered when discussing the greatest of all time by measuring each player with advanced statistics and letting the algorithm clustered them based off those statistics. Originally Gretzky was seen as the only possible candidate due to his extensive records and impressive career base statistics. But using the combination of advanced statistics with a k-means clustering algorithm, we were able to extend the candidates list to include Wayne Gretzky, Mario Lemieux, Bobby Orr, and Mike Bossy. Current great NHL players like Sidney Crosby and Connor McDavid did not make the final cut in either test. This goes to show how dominate goal scoring was back in the day and how goalies/defenses have improved drastically over the years. Connor McDavid could be considered an honorable mention because he is only 0.03 points per game behind Bobby Orr as of right now, but McDavid still has much left of his career yet to play.