Craig Hatfield, Michael Natola, Brandon Gillis, Angantyr Gautson, Noah Folib
Joanna Blanchard
Foundations of Data Science
25 November 2020

Final Project

Part I Executive Summary:

During our project we had encountered many new things about hockey, the NHL, and every player. Using all of the tips and tricks we have learned this year in a very fun and interactive data science class, we created a very good project. In our project we tried to create a team of the most balanced 18 players(3 centers, 3 left wings, 3 right wings, 6 defensemen, and 2 goalies) of all time. Also, we found the best coach to lead this team. Most importantly we did not find statistically the best NHL players of all time, but instead found the most balanced players that could make this team unstoppable.

In question 1, we found out which players scored most goals and got the most assists in their average season, this helped us find out what attackers contributed to most points and could get into our team based on their ability to score goals and assist them. This code showed us that Wayne Gretzky contributed most points on average per season, that was not enough to make the team though as we didn't choose our team only based on goal contributions. We also found out that Mike Bossy was the best goalscorer on average per season, which helped him get into our team. Question 2 had two different parts(a and b) where we found which defenseman scored the most points and which player had the best plus/minus. Having played a very short career due to injury Bobby Orr was amazing during his time. Averaging more points and goals than any other defensemen in NHL history. We also found that Larry Robinson had the best plus/minus of all defensemen but offensively he was not good enough to make the team and Denis Potvin made it over him. In question 3, we found the best players that performed in the post-season or the playoffs. Bringing up Mike Bossy again we found he was amazing at scoring in the playoffs averaging 7 more goals than the next highest. In question 4a, we found what player played the best on the powerplay. One of the best goal scorers in history Alex Ovechkin averaged more power play goals than any other player and also leads in points. He makes a great edition to the team and can be a problem for many teams. In question 4b, we found the best player on the penalty kill. This is very important because if our team takes any penalties we should have the best to handle this situation. Mike Richards averaged 1 more point on the penalty kill than any other player. That is very impressive but still not good enough to make the team. In question 5, we found out what coach we wanted in our team, we decided to choose him by average regular season and postseason wins. The results were that Todd McLellan had the most regular wins on average per season and had one of the most wins in postseason on average, which is why we decided to make him our coach. Question 6 had two parts(a and b) where we found the goalies for our team. In part a, we found that Ken Dryden had averaged the most wins in his career. This was good enough to make the team and he is one of our goalies. In part b, we found that Dominik Hasek had the best save percentage throughout his career saving more than 92 percent of the shots. That is very impressive and was good enough to make the team. In question 7, we found out which players took the least amount of penalties. Val Fonteyne only took 26 penalty minutes in 820 games in his career. Most players take more than that in just one year. He is one of the best disciplined players ever, but he did not make our team. In question 8, we found the player who won the most trophies. These trophies include the Stanley Cup, the Conn Smythe, the

Hart Memorial, the Calder, the Vezina and many more. Not surprisingly, Wayne Gretzky leads this category with 49 trophies. We all know he is the best player of all time and this is just more to add to his resume. In question 9, we found the average age of the players in their prime. The prime meaning when they are at their highest athletic performance and are producing at their best. We found players played the best between the ages of 23-27. In our final question, we found the most balanced players to put on our team that would lead us to success. Using many lines of code we created a team with 18 players and a coach.

Part II Data & Approach:

Our data comes from the website Kaggle in the form of a CSV file. It contains data from all 31 NHL teams and all of the players and coaches involved. This data was collected starting from 1909 to 2011. It has many different sections of the data. We did use Master.csv [7761,31], AwardsPlayers.csv [2091,6], Scoring.csv [45967,31], Coaches.csv [1812,14], and Goalies.csv [4278,23]. We decided to select certain sections of the data because the whole data set has a lot of files, rows, and columns. This will make the data set easier to manage. We dropped all missing data with no first name, no last name, no height, no weight, etc. This will help us because we do not want any NA's in our data.
https://www.kaggle.com/open-source-sports/professional-hockey-database

We assigned the "Scoring.csv" file to scoring_data. This was the most widely used datasets. Out of the 31 columns we used 18 columns and only used the players who were in the NHL (lgID = NHL). This data set would be assigned to a new variable for problems 1-4 and 7.

We assigned the "Coaches.csv" file to coaches_data. This was not used except for one time in question 5. Out of the 14 columns we only used 5 of the columns.

We assigned the "Master.csv" file to master_data. This was not commonly used but a transformed data set was used in each question. Out of the 31 columns we only used 8 columns of data. We also called master_data for question 9 to be able to find the ages of the players.

We assigned the "Goalies.csv" file to goalies_data_file. Then assigned that to goalies data after making all the na's into 0 to help with the calculations later on. Out of the 23 columns we only used 8.

We assigned the "AwardsPlayers.csv" file to awards_player_data. We only used this once in question 7. Out of the 6 columns we only used 2 columns.

Before any of the questions we created two new data sets from master data. We created a"Players" dataset and a "Coaches" dataset. This would be used in every problem and be mutated to be able to display the results found in each problem with their name rather than their playerID. At the end of each problem we would use the left_join command to replace the playerIDs with names.

Question 1: We assigned Scoring_data to d1ID. We grouped the data by playerID and found the average Goals, Assists, and Points using a summarize command. We then displayed the data as d1 and made a graph of the points. We used na.omit to help with calculations.

Question 2a: We assigned Scoring_data to d2aID. We grouped the data by playerID and found the average Goals, Assists, and Points for only the defensemen using a summarize command. We then displayed the data and made a graph of the points.

Question 2b: We assigned Scoring_data to d2bID and grouped by playerID. We then summarized the total plus/minus for each player. We displayed the data as d2b.

Question 3: We assigned Scoring_data to d3ID. We grouped the data by playerID and found the average Goals, Assists, and Points in the average postseason using a summarize command. We then displayed the data and made a graph of the points. We used na.omit to help with calculations.

Question 4a: We assigned Scoring_data to d4aID. We grouped the data by playerID and found the average Goals, Assists, and Points during a power play using a summarize command. We then displayed the data and made a graph of the points. We used na.omit to help with calculations.

Question 4b: We assigned Scoring_data to d4bID. We grouped the data by playerID and found the average Goals, Assists, and Points during a penalty kill using a summarize command. We then displayed the data and made a graph of the points. We used na.omit to help with calculations.

Question 5: We assigned coaches_data to d5ID. We grouped the data by playerID and found the average Goals, Assists, and Points using a summarize command. We then displayed the data and made a graph of the points. We used na.omit to help with calculations.

Question 6a: We assigned goalies data to d6aID and found the average seasonal wins, post seasonal wins and all time wins. We then displayed the data and made a graph.

Question 6b. We assigned goalies data to dbaID and found the amount of games played and the save percent using two other variables in a summarize.We then displayed the table.

Question 7: We assigned scoring data to d7ID and found the games played and the penalty minutes. We then displayed the table

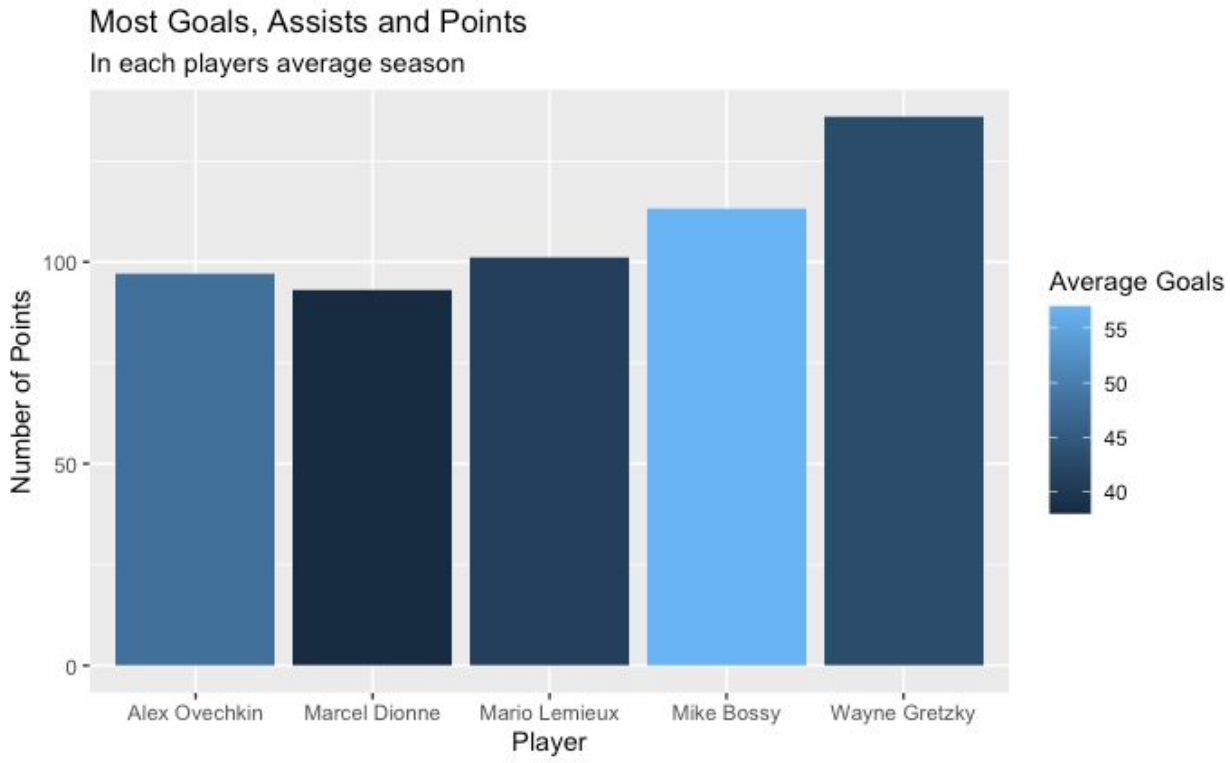
Question 8: We used the awards_players_data and grouped them by playerID. We then summed up the amount of awards each player got then displayed the table.

Question 9: We used master_data and assigned this to d9ID. We used this to find the ages of the players in the middle of their career and find the 3rd quartile to be able to select the players who have not surpassed the top 75% ages.

Question 10: This question ties in every question above and creates a new table using full_joins, left_joins and a lot of array manipulation. We multiple new datasets here simple to help organize the players by their position and then we created the team roster as d10.

Part III Detailed Findings:

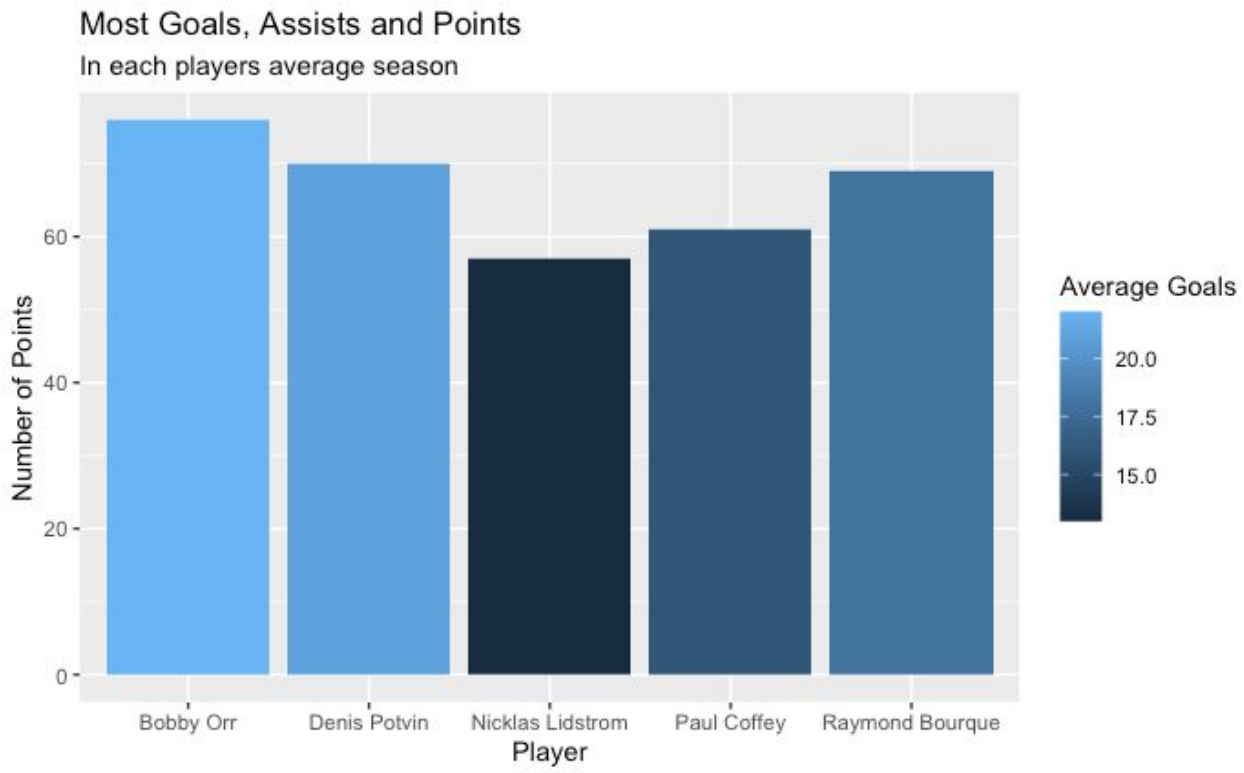
Question 1, What player has the most goals,assists and points in their average season?



We used the data set scoring.csv to get the player's scoring. Out of all the other scoring statistics we chose goals, assists, points. We thought these columns would best show the player's success on the ice.

This graph shows what players had most goals, assists and points on average over their career. We wanted to see what players were scoring most points on average per season in their careers, this helped us select attackers for our team. As we could have guessed Wayne Gretzky has a good lead in 1st place, we found out that Gretzky is not the best goal scorer even though he has the most points, the best goalscorer is Mike Bossy with over 55 goals per season.

Question 2a, What defenseman scored the most points in their average season?



We used the data set scoring.csv to find out how many goals, assists and points defenders had. We felt like these stats would really help us find defenders for our team as we wanted at least one defender who could score goals and assist.

This graph showed us what defenders had the most goals and assists on average over their careers. We feel like it is important for defenders in hockey to contribute well in the offense also, that is why we wanted to see what defenders were getting points. The graph shows that Bobby Orr has the most points on average for a season as well as averaging most goals per season, while Denis Potvin and Raymond Bourque come really close to him in average points per season. This information helped Bobby Orr get into our team as one of the defenders.

Question 2b, What experienced defensemen has the best plus/minus?

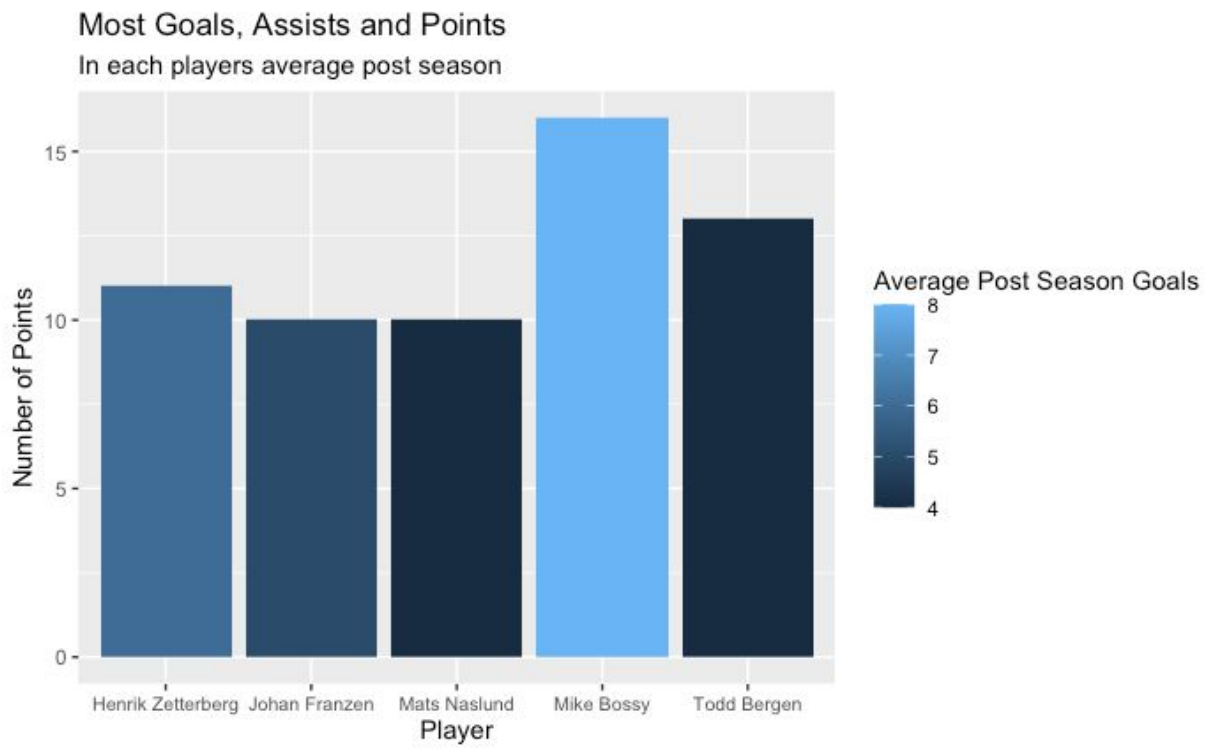
Name <chr>	+/- <dbl>	Games Played <dbl>
Larry Robinson	730	1384
Raymond Bourque	528	1612
Denis Potvin	460	1060
Serge Savard	460	1038
Nicklas Lidstrom	450	1564
Brad McCrimmon	444	1222
Scott Stevens	393	1635
Mark Howe	390	866
Al MacInnis	373	1416
Brad Park	358	1113

1–10 of 20 rows Previous 1 2 Next

We used the data set scoring.csv. When trying to find the top defenders based on there +/- this allowed us to be able to pinpoint the best defenders in the NHL. The +/- tells us which plates make the team better as a whole when they were on the ice. This gave us insight into which players didn't just have the crazy stats but were good team players.

This table shows what defensemen have the best +/- when playing and we wanted them to be experienced to get a better view over it so we only looked at players that have played at least 750 games. When players have good +/- stat it shows that they are playing well when they are on the field but it could also just be that they are on the best team, with that we could not just choose players based on their +/- stat, this table helped us though and Denis Potvin who is number 3-4 in the table was selected in our team.

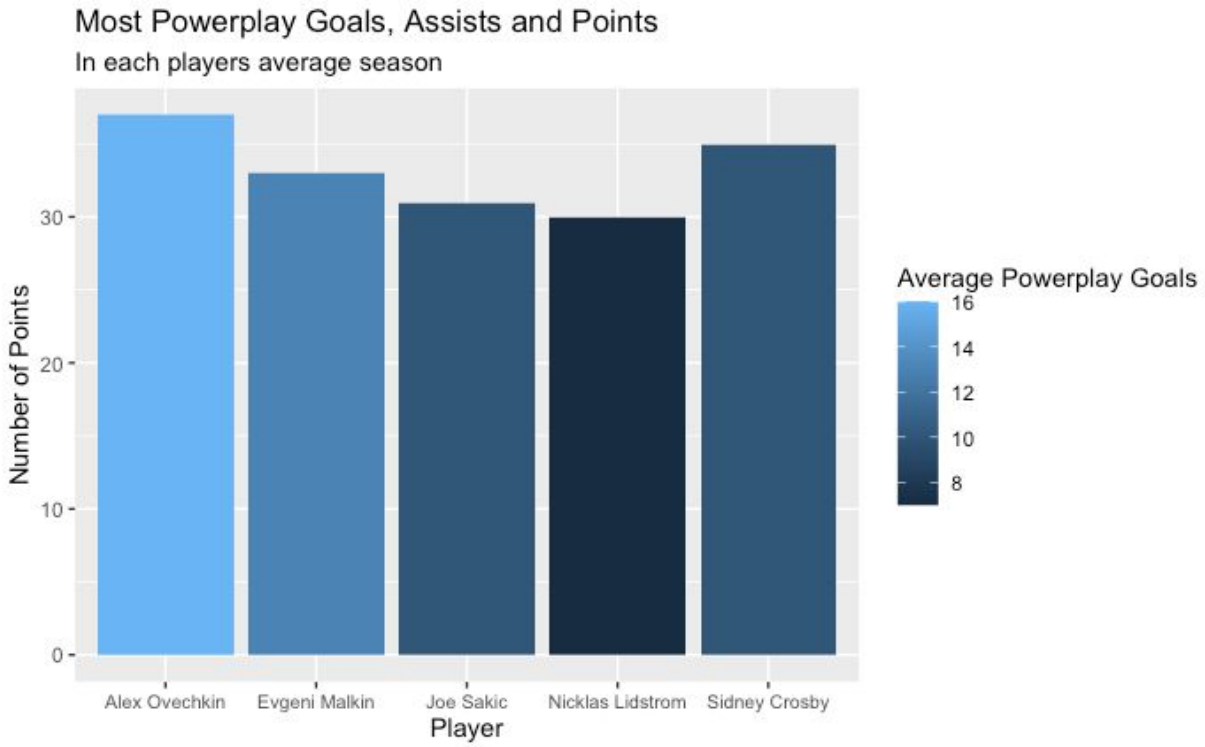
Question 3, What player has the most goals,assists and points in their average postseason?



We used the data set scoring.csv for this question. We chose the columns PSG(post-season goals) and points for this question.

This graph shows what players score the most points on average in the postseason. We wanted to see if there were any players doing really well in the postseason, because it is harder to play in postseason and therefore we want players that can handle the pressure and still score goals and assist. It is not shocking to see that Mike Bossy has the most average points and goals in postseason as he was far ahead of every other player in normal season goals. This tells us that Mike Bossy is arguably the best goalscorer in NHL history. We can also see in this graph that Todd Bergen scores a lot of points in the postseason and that most of those points are assists.

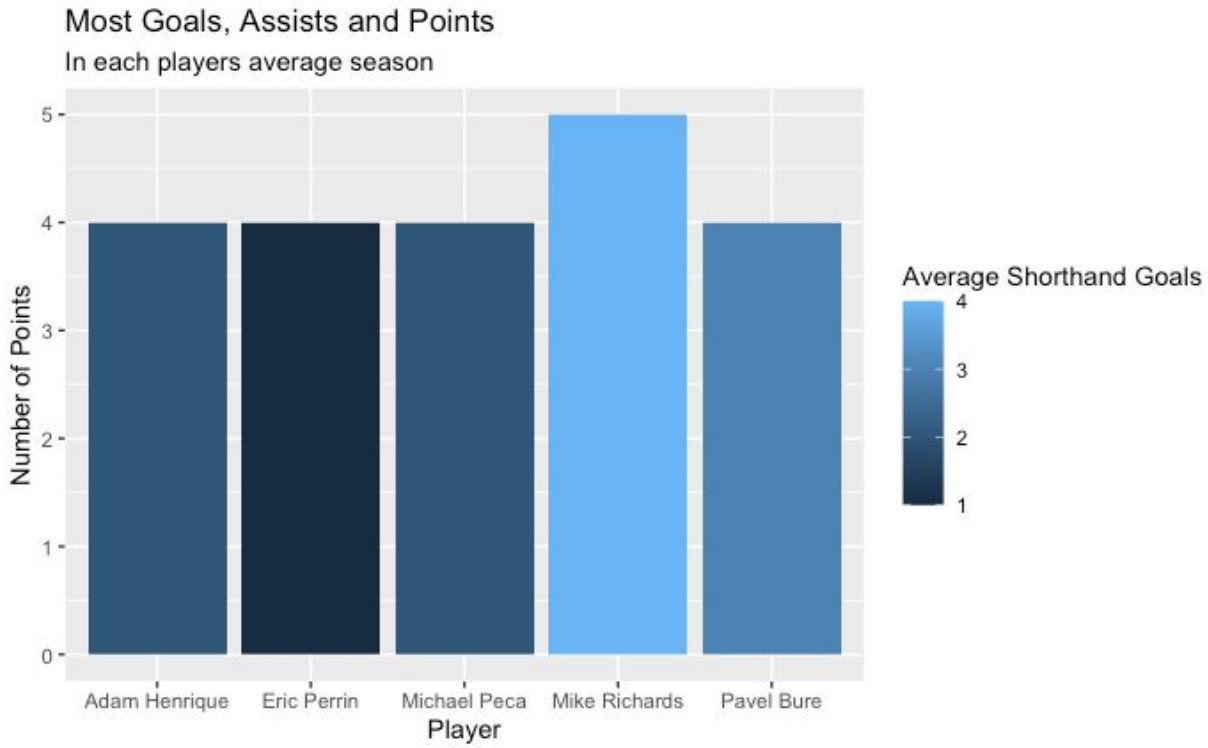
Question 4a, What player has the most goals,assists and points in their average power play per season?



We used the data set scoring.csv to find out which players did best during powerplay, we decided to look at the columns Goals, Assists and Points as we felt like they would tell us what players played best during powerplay.

This graph shows what players do best in powerplay, we wanted to see that because powerplay is an important part of hockey. As the graph shows Alex Ovechkin is the best Goalscorer when he is in powerplay as well as scoring most points total on average per season. Sidney Crosby is not far behind him in points but is not scoring as many goals. Malkin, Sakic and Lidstrom are not far behind in points scored and none of them are scoring as much as Ovechkin either.

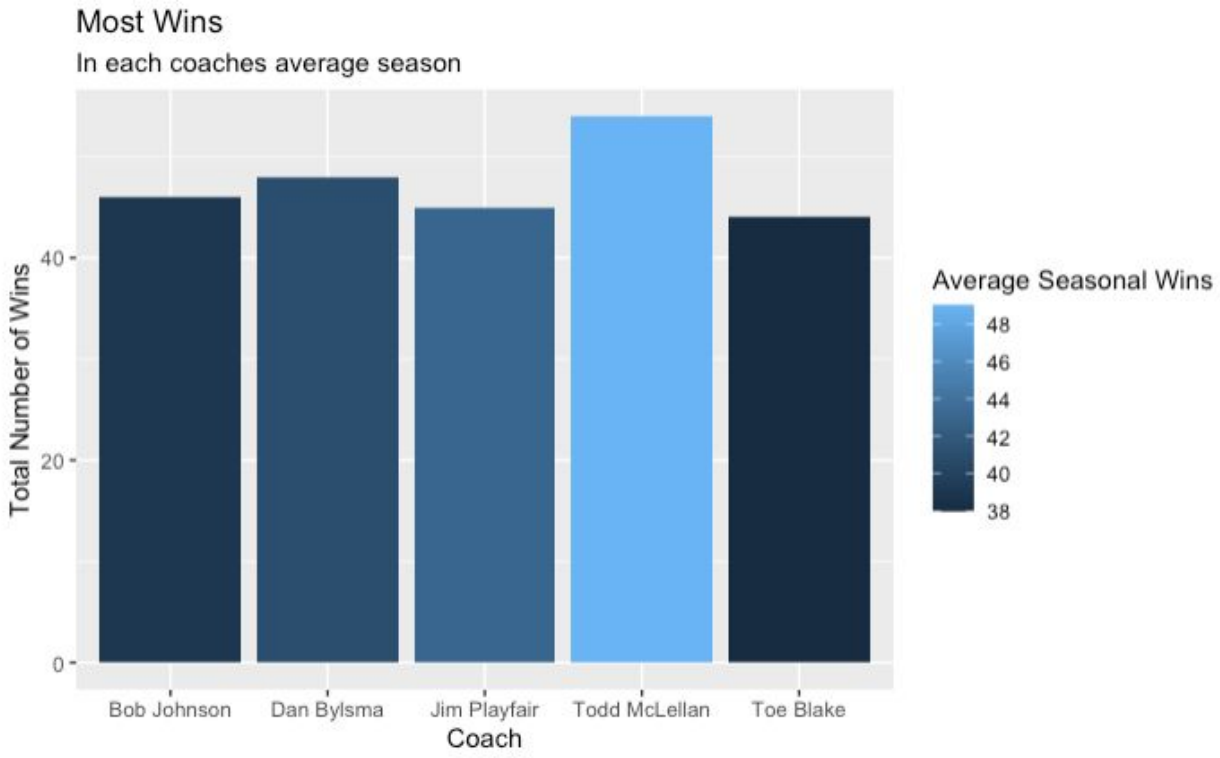
Question 4b, What player has the most goals,assists and points in their average Penalty Kill per season?



We used the data set scoring.csv for this question. We chose the columns SHG(short-handed goals) and points for this question. After we looked at what players did best in powerplay, we also wanted to see what players did best on average penalty kill.

The first thing that we noticed was that none of the top 5 players in the powerplay graph were in this graph, and also as we knew before the average points scored in this graph is much lower than in the other one. The results showed that players don't score many points short handed and only Mike Richards has more than 4 points on average per season, coming after his 5 points there are 5 players trailing him with 4 points on average per season. We can also see that Mike Richards is the best goalscorer when playing short handed.

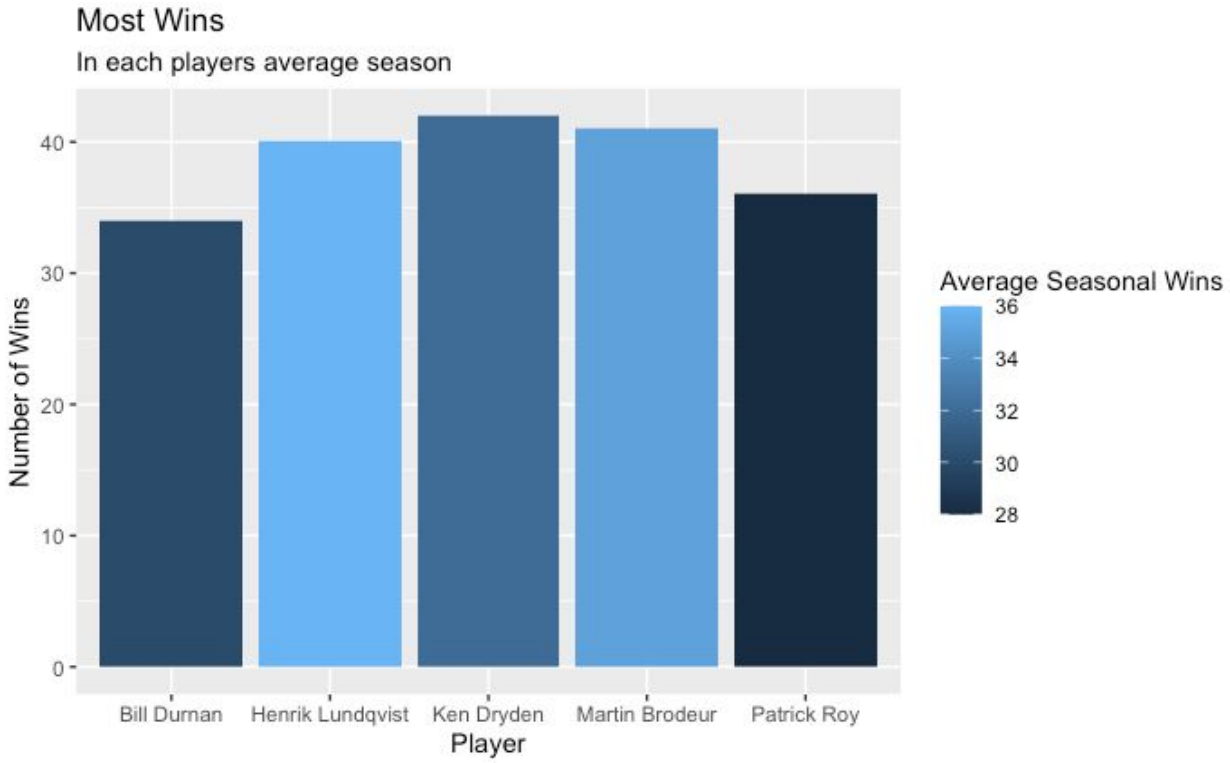
Question 5, What coaches have the most wins in their average season, post season and all time?



This graph shows us what coaches have the most wins on average per season in their careers. The coaches.csv dataset was used here to help us create this graph. Using the information we found from the dataset we were able to mutate wins,postseason wins and all time wins to find out which coaches had the best average season.

We needed to pick a coach for our team and we found it a good idea to see what coaches were winning most games on average both in the normal season and postseason. The graph shows that Todd McLellan is averaging most wins per season but Dan Bylsma is not far behind and he has more postseason wins on average also.

Question 6a, What goalie has the most wins in their average season, post season and all time?



This graph shows what goalies have the most wins on average per season in both regular and postseason. We used the Goalies.csv data set to find the information that we needed to create this graph. Goalie is a very important position in hockey and we knew we had to be careful choosing our goalies and as the graph shows Ken Dryden has the most wins on average in both regular- and postseason, Henrik Lundqvist and Martin Brodeur come the closest to him but as the colors show Dryden is better in postseason which is why he is more ahead of the other ones in our opinion than the graph shows.

Question 6b, What goalie has the best save percentage in their average season, post season and all time?

Name <chr>	Games Played <dbl>	Save Percent (%) <dbl>
Dominik Hasek	735	92.23
Roberto Luongo	727	91.93
Tomas Vokoun	680	91.68
Miikka Kiprusoff	599	91.36
Martin Brodeur	1191	91.30
Jean-Sebastien Giguere	557	91.30
Evgeni Nabokov	605	91.24
Patrick Roy	1029	91.02
Marty Turco	543	90.96
Jose Theodore	633	90.94

1–10 of 20 rows Previous 1 2 Next

This table shows what goalies have the best average save percentage in their average regular- and postseason. We chose the column names playerID(name), GP(games played), and save percentage. We used the information we got from the goalies.csv data set to help us mutate the save percentage into its own column since it was not already a column.

We knew it was not enough to just see how many wins goalies were averaging per season so we wanted to see their save percentage also, to be on the table goalies had to have played at least 500 games. We noticed that Martin Brodeur who was in top 3 in the last graph made it to top 5 here, which makes him a very good goalie. Dominik Hasek has the best save percentage of all the goalies to have played over 500 games, that makes him a really strong candidate for our team also.

Question 7, What experienced player took the least amount of penalties?

Name <chr>	Games Played <dbl>	Penalty Minutes <dbl>
Val Fonteyne	820	26
Bill Quackenbush	774	95
Woody Dumart	772	99
Butch Goring	1107	102
Dave Keon	1296	117
Robert Kron	771	119
Rick Kehoe	906	120
Don Marshall	1176	127
Phil Goyette	941	131
Mikael Andersson	761	134

We used the data set scoring.csv for this question, we wanted players that would not get too many penalty minutes, which is why we chose the columns “GP” and “PIM” and then we changed the names of the columns to “Games Played” and “Penalty Minutes”.

This table Shows us what players get the least amount of penalty minutes and how many minutes they get, it also shows us how many games they played and we wanted them to have played at least 750 games. We wanted to see if any of the players that were close to getting in our team had fewer penalty minutes than players with similar stats, because if that were so we would prefer to choose a player with a little worse stats and fewer penalty minutes. Fewer penalty minutes gives the team a better chance of winning more games.

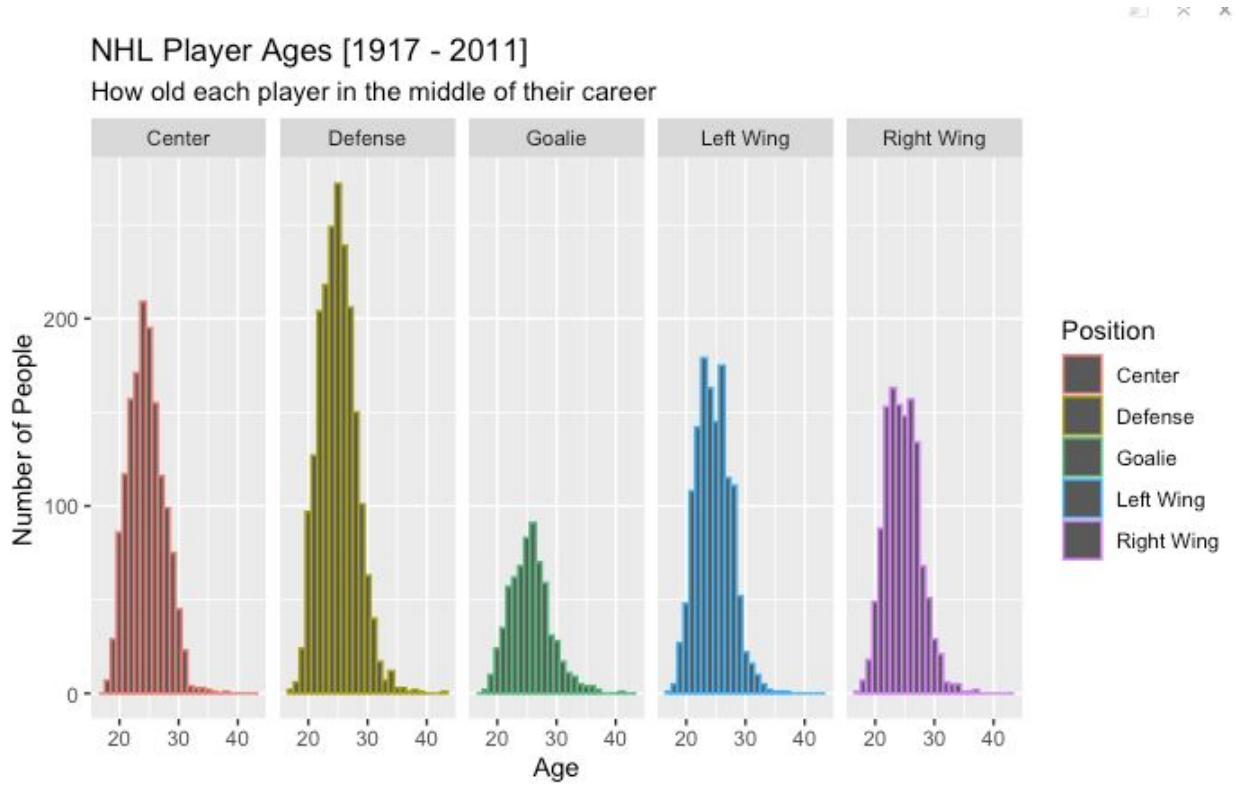
Question 8. Who are the greatest players of all time based off of Awards they received?

Name <chr>	Number of Awards <int>
Wayne Gretzky	49
Gordie Howe	33
Mario Lemieux	28
Raymond Bourque	26
Bobby Orr	26
Nicklas Lidstrom	21
Dominik Hasek	20
Martin Brodeur	18
Doug Harvey	18
Bobby Hull	18

1–10 of 50 rows Previous 1 2 3 4 5 Next

We used the data set awards.csv for this question. We chose the columns playerID(name) and we counted up the number of awards each player got using the sum() command. This table shows how many awards the players got during their careers. Awards for individual performance really show how good players actually are, that is why we wanted to add that into our research on players for our team. It was not shocking to see Wayne Gretsky number 1 in awards received with a dominant lead on the others. It was not enough to only get many awards to get into our team though as we wanted a balanced team with great players, but it definitely helped to have many awards to get into our team roster.

Question 9, What are the ages of NHL players in the middle of their career? Find the standard deviation or third quartile for each position.



We used the data set master.csv for this question. We chose to separate the graphs into 5 sections for each position using the command facet_wrap(). This helped us look at the data better and the colors used were also very helpful. This graph shows how old players are in the middle of their careers by position. We wanted to find out how old players in the NHL were in the prime of their careers on average by each position. This helped us find out what was the average age of players in their prime by each position and we used that when we selected our team, we wanted to have players in their prime and get rid of players who were too old in our opinion. The graph shows that most players were between 23-27 old in their prime.

Question 10. Who would we want on our team?

Name <chr>	Position <chr>
Evgeni Malkin	Center
Nicklas Backstrom	Center
Anze Kopitar	Center
Alex Ovechkin	Left Wing
Henrik Zetterberg	Left Wing
Ilya Kovalchuk	Left Wing
Mike Bossy	Right Wing
Theoren Fleury	Right Wing
Dany Heatley	Right Wing
Denis Potvin	Defense
Bobby Orr	Defense
Scott Niedermayer	Defense
Behn Wilson	Defense
Dion Phaneuf	Defense
John-Michael Liles	Defense
Ken Dryden	Goalie
Dominik Hasek	Goalie
Todd McLellan	Coach

We used all of the data sets for this question. Combining all of the data sets into one question was hard but doing this gave us our team. This table is who our code told us we should put on our team roster. We found this by taking each question above and seeing what players were top 50 in each category and which players were top 50 multiple times. We summed up the amount of times a player was top 50 and kept the players with the most top 50, but we had to eliminate players above 27 year old because they would be 75% done with their career (found in question 9). So for example Nickolas Lidstorm was an amazing defenseman with 5 top 50 placements but he was 31 years old. These are the most balanced players according to our code.

Part IV Validity & Reliability Assessment:

Our project was pretty accurate and represented our data well. Looking at our analysis compared to others shows that we correctly managed to find some of the most balanced 6 players to ever play hockey plus a coach. As a group, we talked to 2 Division 1 hockey players that play for Merrimack. They are obviously great hockey fans and validated our results saying “these players are some of the best to ever play and their games are very balanced”. The players we picked impacted the game positively and many new young players looked up to them for advice and encouragement.

Another factor that helped verify our data was the dataset “AwardsPlayers.csv”. If a player had gotten an award they were automatically boosted in the ranking to determine who we were going to choose. This would help merge who we, as a team, would decide to who are the best players and who the officials thought were the best players as well.

Part V Appendix(our code):

On a different document*