Data Story - Capstone Project Milestone Report

Molly McNamara

Hockey and the Potential Impact of Data Science

The National Hockey League (NHL) is considered to be one of the four major professional sports leagues in the United States (though ranked behind baseball, football and basketball). The NHL is a binational entity split between the US and Canada, where hockey is by far the most popular sport. The economics of hockey are complicated by the shared border and exchange rates, but team success ultimately drives financial success for the team and the league. Teams see an impact of improved performance and win totals on season ticket sales and sponsorship revenue.

Establishing the right balance of player skills and team needs to make a successful run at the playoffs is the challenge of every NHL general manager (GM). While in-game coaching decisions are left to the coaching staff, the GM is responsible for acquiring and dismissing players as well as the coach. A skilled GM evaluates the existing team members and identifies gaps or needs in order to find and obtain talent that will support and improve the team as a whole. Better analysis of player and league data can help guide a GM in improved decision making when acquiring team talent, hopefully resulting in a more successful team.

The Data

The project dataset consists of NHL player, team and coaching statistics for years 2000-2011 and NHL draft data from the late 1980s through 2010. These datasets include player names, ages and birthdates, countries of origin, positions played, scoring and defensive statistics by year, junior level and NHL teams and leagues played in, coaching records, and more. The information is gathered from several sources and in several separate files and formats.

Data Wrangling

The datasets for this project were obtained from The Hockey Database and Hockey Reference. Data from The Hockey Database was available in CSV files by subject; data from Hockey Reference required download, compilation in a TXT file and conversion of the entire dataset to CSV.

Four files were combined to create the final cleaned dataset for analysis: Master (basic demographic information on all NHL players from 1908 to 2011), Scoring (scoring statistics for all NHL players from 1908 to 2011), Teams (team statistics and playoff outcomes from 1908 to 2011), and Draft (player draft information from 1979 to 2010). The date range for draft data was determined by identifying the earliest drafted player still playing in the period of interest for this project (the project will evaluate game data from 2000-2011 - a player drafted in 1979 was still playing in 2000) and the latest possible drafted player who would play in this time period (2010). Note that there is no data for 2004 due to a league-wide labor dispute (lockout) wherein the entire season was cancelled and no games were played.

The R packages used for data wrangling were as follows:

```
install.packages("dplyr")
install.packages("readr")
install.packages("tidyr")
install.packages("countrycode")
```

Columns for all 4 files were renamed to understand the variables more easily. Columns for player name-ID and amateur team-league were split into their separate components.

Unnecessary or duplicative variables were removed. Multiple fields were converted to factors or integers as appropriate. Birth and death month/day/year columns were joined into a single date column for each event. The datasets were subset to the years 2000 to 2011.

The datasets were evaluated with the summary() command to assess the presence of any obvious outliers. NA values in both sets were identified. In the case of certain variables, such as death statistics, NA is appropriate as most of the players in the subset timeframe had not yet died. Scoring statistics that displayed NA incorrectly (for example, a player did not score any goals in a given year) were replaced with a 0.

New variables were created for team playoff success, birth regions, draft round, and additional scoring statistics. Finally, the cleaned datasets were merged by year and team and the new cleaned data file was written and saved as CSV file.

The data wrangling report can be found here: https://github.com/bouncebarkrun/SpringboardCapstoneProject/blob/master/DataWranglingReport.Rmd

Final Dataset Composition

- Team ID, Conference and Division
- Team Season end rank and Playoff result
- Team made playoffs (0 or 1)
- Team Games, Wins, Losses, Ties, Overtime Losses, Points, Shootout Wins and Shootout Losses
- Team Goals For and Goals Against
- Team Penalty Minutes and Bench Minors
- Year (Season)
- Player ID
- First Name and Last Name
- Height and Weight
- Shooting Hand
- First and Last NHL season
- Years of Experience Position Played
- Birth Country, Birth State/Province, Birth City
- Death Country, Death State/Province, Death City
- Birthdate and Deathdate
- Birth Region
- Stint with team
- Games Played
- Goals, Assists and Points
- Penalty Minutes
- Plus Minus Rating Powerplay Goals and Assists
- Shorthanded Goals and Assists
- Gamewinning Goals and Gametying Goals
- Shots
- Goals and Shots Per Game
- Percent of Team Goals Scored
- Percent of Team Games Played
- Draft Pick and Draft Round
- Draft Year
- Draft Team
- Draft Age
- Amateur Team and League

Limitations

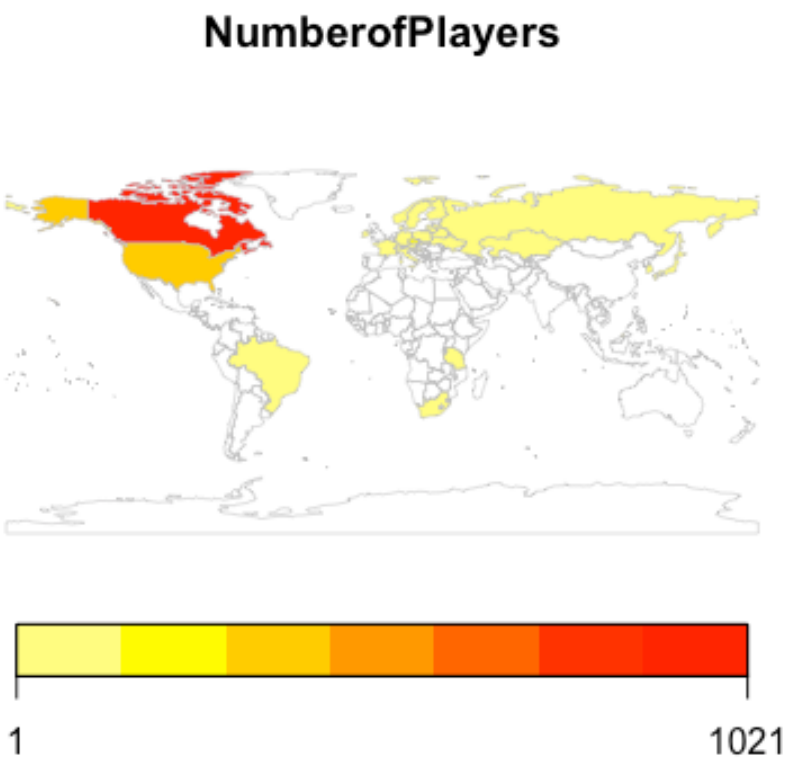
Several limitations can be identified in the final dataset at this point: (a) the lack of play statistics prior to draft and draft combine results (the latter is kept confidential by the NHL), which could have been useful in predicting future NHL performance and team success; (b) this data largely

ignores the defensive contribution of goaltenders by omission of goaltending statistics; and (c) Time On Ice (TOI), a well-known parameter of player performance, is not available.

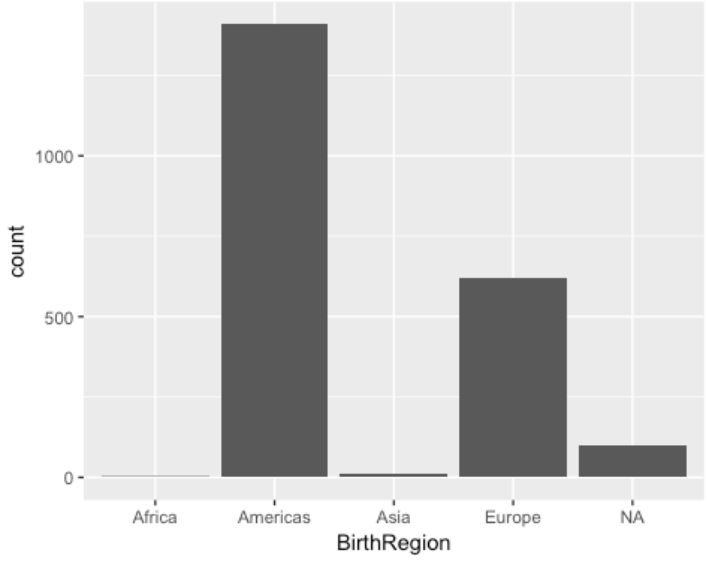
Preliminary Exploration

The purpose of the initial data exploration was to screen for any obvious trends in the data and evaluate the relationships between player demographics and characteristics and teams making the playoffs.

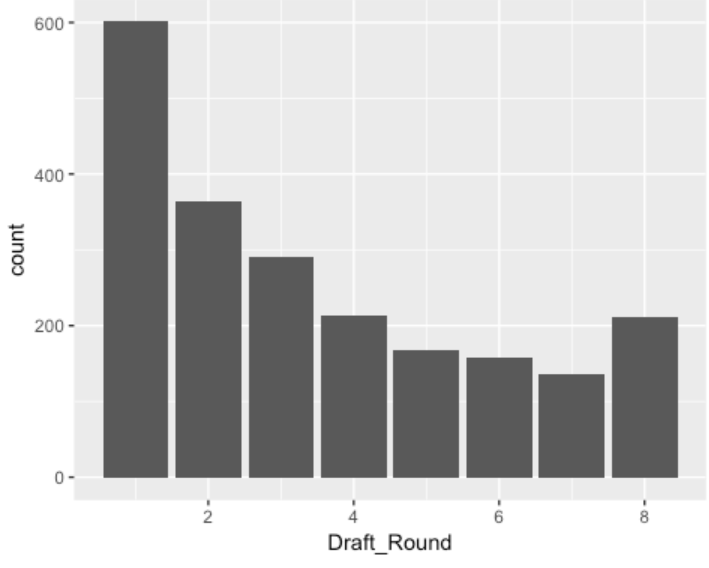
Player birth countries were mapped to a world map to evaluate the many origins of players in this dataset.



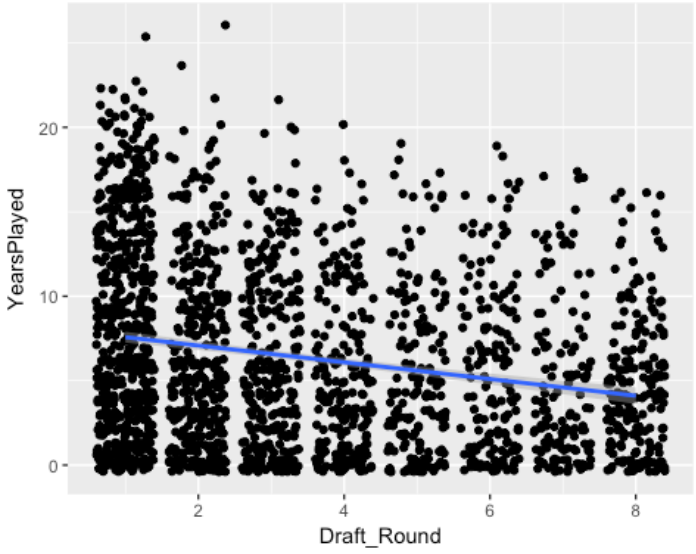
While it appears drafted players originate from many different countries, the NHL is based in the US and Canada; it might be useful to see regionally where the most players come from. Based on the distribution below, the Americas and Europe generate the vast majority of players drafted in the NHL.



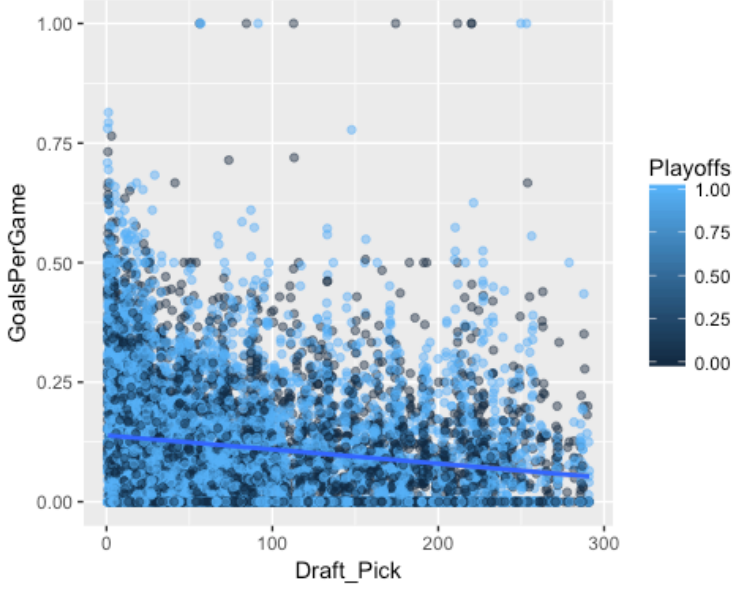
Draft statistics were initially evaluated by viewing a distribution of the draft rounds of the players in this dataset.



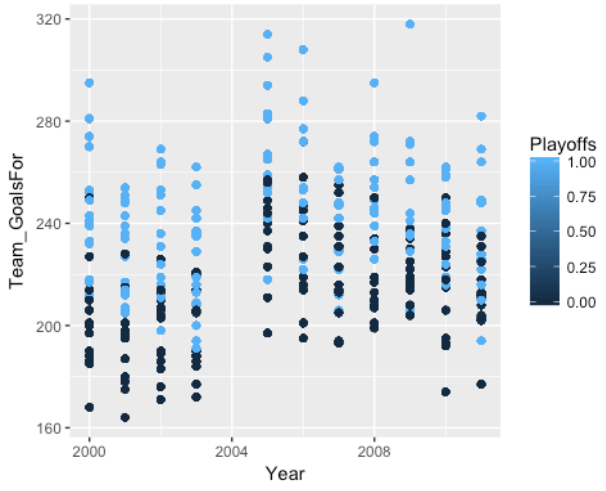
The draft round in the dataset appears to be skewed towards earlier rounds. A comparison between draft round and active years (up to 2012) in the NHL supports the idea that players from earlier draft rounds may have somewhat longer NHL careers.

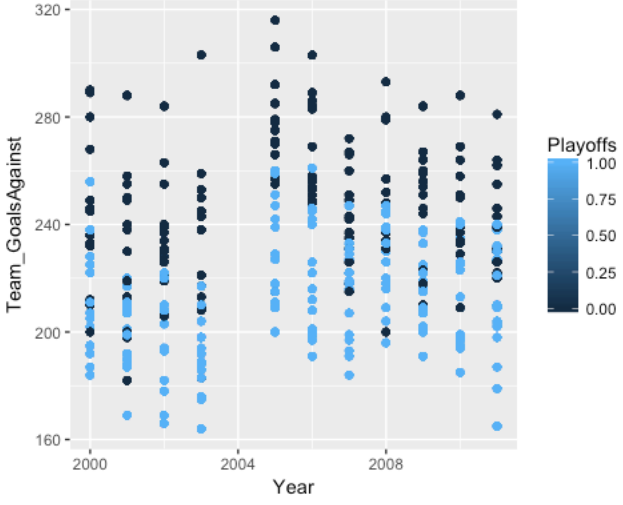


Another way to evaluate relationship between player draft position and team playoff success might be to see if higher-drafted players score more goals per game. It does look like there is a slight trend for more goals per game from higher-drafted players but it's not clear if this impacts making the playoffs.



It should be expected that teams who score more goals in a year and allows less goals in a year are more likely to go to the playoffs.





This hypothesis appears to be true for the highest goal-scoring teams and lowest goal-allowing teams year-to-year.

The initial data exploration report can be found here: https://github.com/bouncebarkrun/SpringboardCapstoneProject/blob/master/StatisticalAnalysisReport.pdf

Modeling Approach

In examining the data in more detail over the course of the wrangling and exploring exercises, several simple truths are clear: teams that win more games, score more goals and allow less goals by their opposition go to the playoffs. However, as teams are built and restructured over years, the question still exists of what player characteristics contribute most strongly to success. The main goal of the project remains the primary focus in the project proposal - to determine what factors contribute to regular season success that is realized in a playoff berth.

The project approach at this point is to:
* Build the best predictive model of team success (defined as making the playoffs) from the player and draft characteristics.
* A secondary aim will be to define underrated and overrated players, determined by draft position and best 3 years of career performance, and evaluate if player performance at the time of the draft is predictable by amateur team, amateur league, nationality, position played or other factors.