Data Collection and Visualization Project

Dylan Lancaster

GOALS

This project utilizes BeautifulSoup to scrape desired NHL player attributes that are acquired during the 2020-2021 NHL season. The attributes are collected from QuantHockey. After writing the collected attributes to CSV files, we can utilize MatPlotLib to visualize the data we have retrieved, namely correlations between:

Average Time on Ice and Points Earned by Forwards

Total Number of Hits and Total Penalty Minutes Earned by Defensemen

Age in Years and Save Percentage Earned by Goaltenders

DESCRIPTION OF DATA

There are three .csv files within the project directory containing player data for their respective role.

The details of the .csv files are as follows:

forward_data.csv contains individual player data for all active NHL players for the 2020-2021 season who are either Centers, Left Wingers, or Right Wingers. By default, the entries are sorted by points earned in descending order.

Details on the column data:

Last name: The last name of the player

First Name: The first name of the player

Age: The age of the player in years

Height: The height of the player (in ft/in)

Weight: The weight of the player (in lbs)

Team: The name of the NHL team the player is on

Abbr: Three-letter abbreviation of the 'Team' attribute

GP: Total number of games played by the player

G: Total number of goals scored by the player

A: Total number of assists earned by the player

P: Total number of points earned by the player

TOI: Average time spent on the ice per game by the player

Hits: Total number of legal hits made by the player

PIM: Total number of penalty minutes earned by the player

+/-: Calculated plus-minus earned by the player (A player is awarded a "plus" each time he is on the ice when his team scores an even-strength or shorthanded goal. He receives a "minus" if he is on the ice for an even-strength or shorthanded goal scored by the opposing team. Power play or penalty shot goals are excluded. An empty net does not matter for the calculation of plus-minus.

defense_data.csv contains individual player data for all active NHL players for the 2020-2021 season who are either Left Defensemen or Right Defensemen. By default, the entries are sorted by points earned in descending order.

Details on the column data:

Last name: The last name of the player

First Name: The first name of the player

Age: The age of the player in years

Height: The height of the player (in ft/in)

Weight: The weight of the player (in lbs)

Team: The name of the NHL team the player is on

Abbr: Three-letter abbreviation of the 'Team' attribute

GP: Total number of games played by the player

G: Total number of goals scored by the player

A: Total number of assists earned by the player

P: Total number of points earned by the player

TOI: Average time spent on the ice per game by the player

Hits: Total number of legal hits made by the player

PIM: Total number of penalty minutes earned by the player

+/-: Calculated plus-minus earned by the player (A player is awarded a "plus" each time he is on the ice when his team scores an even-strength or shorthanded goal. He receives a "minus" if he is on the ice for an even-strength or shorthanded goal scored by the opposing team. Power play or penalty shot goals are excluded. An empty net does not matter for the calculation of plus-minus.

goaltender_data.csv contains individual player data for all active NHL players for the 2020-2021 season who are Goaltenders. By default, the entries are sorted by total games played in descending order.

Details on the column data:

Last Name: The last name of the player

First Name: The first name of the player

Age: The age of the player in years

Height: The height of the player (in ft/in)

Weight: The weight of the player (in lbs)

Team: The name of the NHL team the player is on

Abbr: Three-letter abbreviation of the 'Team' attribute

GP: Total number of games played by the player

GAA: Average number of goals made against the player

SV%: Calculated save percentage earned by the player

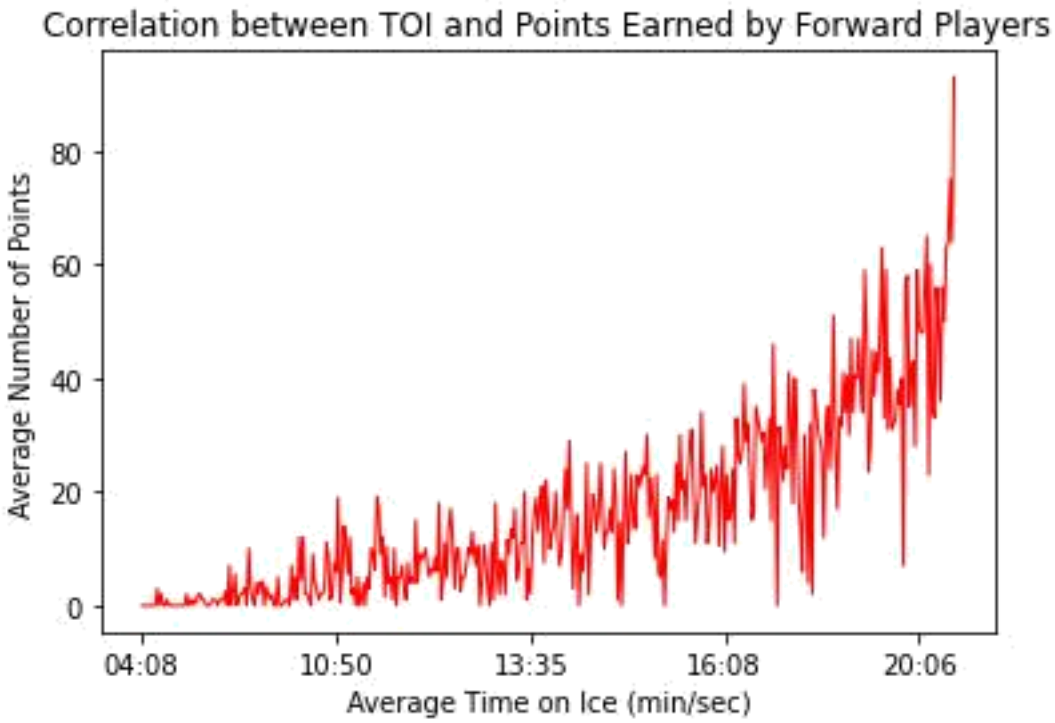
SV: Total number of saves made by the player

SO: Total number of shutouts made by the player

DESCRIPTION OF VISUALS

Visual 1

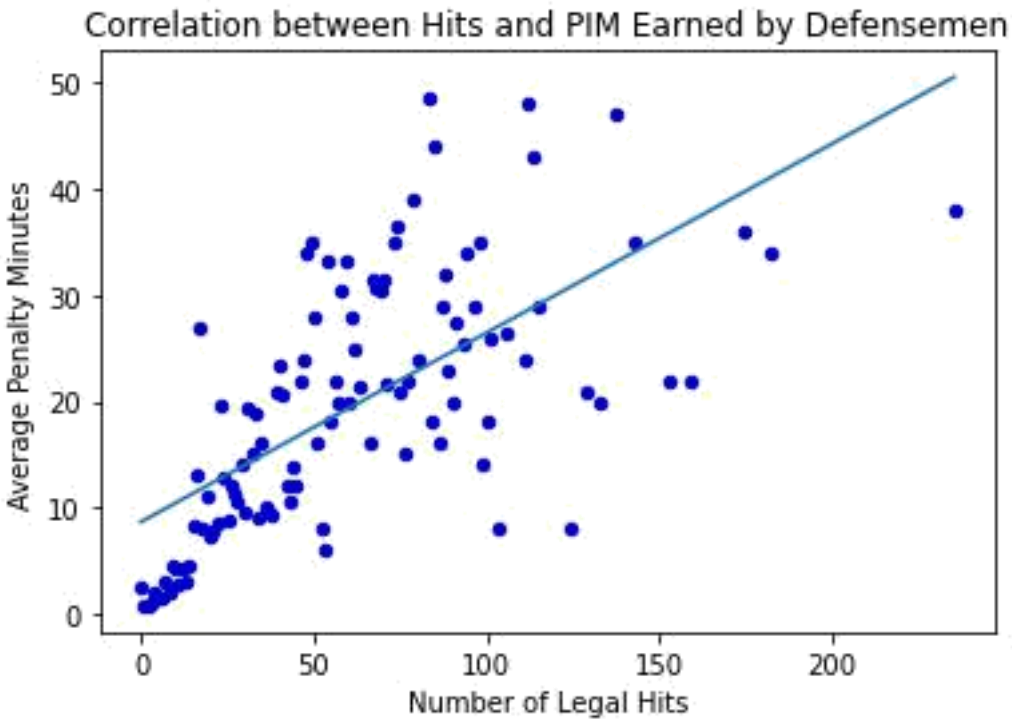
To be an NHL player, you must earn your play time. It would make sense that a player who spends more time on the ice has more opportunities to acquire points (the sum of goals and assists). Within the *forward_toi_points()* function, I created a line plot to show the correlation between average time on ice and total number of points acquired by forward players from the *forward_data.csv* file.



Based on this visual, we can easily see a positive correlation between an increase in average time on ice leading to an increase in the total number of points earned by each player. The curve on the graph looks almost exponential in nature, reinforcing the fact that players who spend more time on the ice per game are more valuable towards their team in producing points.

Visual 2

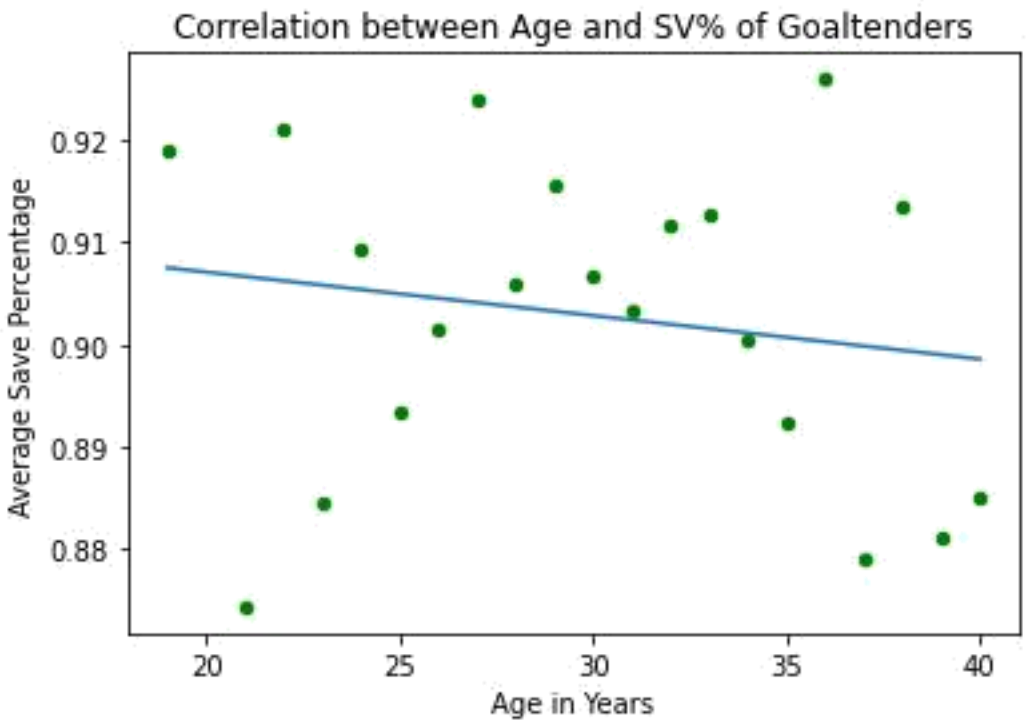
Typically, those who earn more penalty minutes are more aggressive players, and are sometimes known as 'enforcers'. Conveniently enough, the NHL tracks the number of legal hits via a stat known as 'Hits'. I wanted to see if, given a defenseman's track record of hits, if their aggressive playstyles correlated with a longer amount of time spent in the penalty box. Within the *defense_hits_pim()* function, I created a scatter plot to show the correlation between total number of legal hits and total number of penalty minutes earned over the season acquired by defensemen players from the *defense_data.csv* file. I also included the line of best fit for the data.



Based on the strong positive correlation of the plot points, and further supported by the line of best fit, we can clearly see that given a player with a higher number of legal hits, he will more than likely have a higher number of penalty minutes attributed to him. This further supports the initial idea that players who hit hard and often are also those who are more likely to commit infractions.

Visual 3

To be a goaltender in the National Hockey League, you must have lightning-fast reflexes and thrive under pressure. When your forwards lose control of the puck, and your defensemen are unable to make the defensive play necessary, you are all that stands in the way from your opponent scoring a morale-ruining goal. With the natural degradation of the physical body and reflexes that comes with age, I thought it would be interesting to visualize the connection between a goaltender's age and his save percentage. Within the *goalie_age_svpctg()* function, I created a scatter plot to show the correlation between the age in years of an NHL goalie and the average save percentage for all goalies at that age from the data present in *goaltender_data.csv*. I also included the line of best fit for the data.



Based on the data, evidenced further by the line of best fit for the data, there is a weak negative correlation between an increase in an NHL goaltender's age and their average save percentages. While it does not produce a very strong result, this does support our initial hypothesis that younger NHL goalies will typically have higher save percentages than those who are older.