NHL Report: Does Age Matter?

Ben Turner and Aidan Morris

12 March 2022

Introduction

The National Hockey League (NHL) is a collection of the best hockey players from around the world that come together to play in cities across North America. These players range in many different characteristics, including age. It is always fun to watch fresh young faces join the league and make an instant impact. Their energy and determination seems to suggest they will outperform older players that are losing their physical edge. However, it is also incredible to see older famous players continue to impress with their spectacular performances. An interesting question we set forth to answer is whether the younger more energetic NHL players perform better than the older yet more experienced players.

Methods

Our data set was obtained from hockey-reference.com[1], a website that collects data on hockey teams and players. This website includes performance data on all professional National Hockey League (NHL) players in the 2021-2022 season.

The variables that dictate players' performance are Rk (an ID for the players), Age (measured in years), PTS (total number of points)[2], X... (plus/minus)[3], and BLK (total number of blocks by each player). These variables provide the most accurate representation of a player's performance both offensively and defensively. Additionally, the players were divided into one two categories, ages greater than 26, or ages less than or equal to 26[4].

This data set can be used to evaluate the hypothesized relationship between the age of NHL players and their corresponding performance during the 2021-2022 season.

[1] https://www.hockey-reference.com/leagues/NHL_2022_skaters.html

[2] points is the sum of goals and assists for a given player

[3] plus/minus for a player is how many goals your team scores minus how many the other team scores when that player is actively playing on the ice

[4] this was chosen as the division as it is the median age of all players in the data set

Results

This study looked at a total of 956 professional hockey players and their accompanying 2021-2022 season performance statistics.

As seen in Fig. 1, the total point distribution is skewed right, with no obvious outliers and a majority of players having a total point count of 25 or under. The mean number of total points scored for all players was 15.497, with a standard deviation of 16.385. The plus/minus (X...) distribution seen in Fig. 2 has a normal distribution, with a mean of -0.403 and a standard deviation of 9.369. Lastly, Fig. 3 displayes the distribution of total blocks performed by each player appears to be skewed right, with no obvious outliers, a mean of 25.445, and a standard deviation of 27.164.

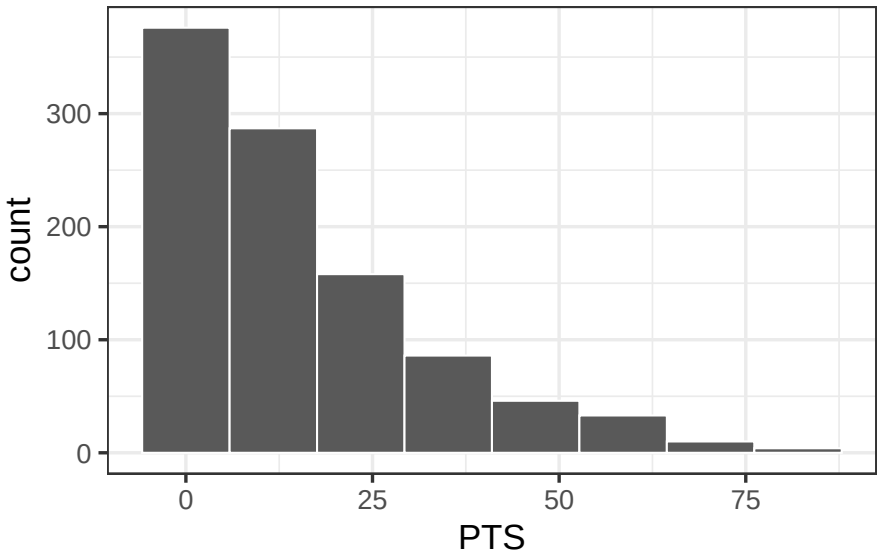

Fig. 1: Number of points for NHL players with 8 bins.

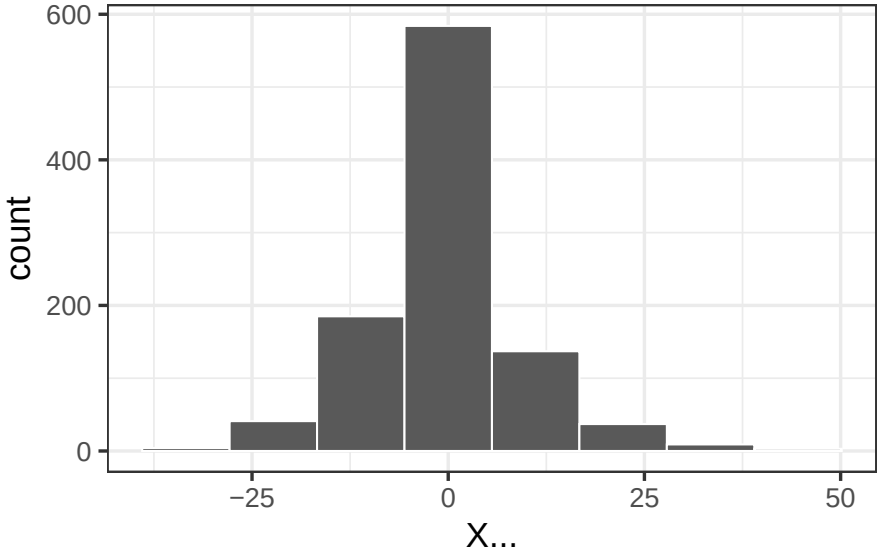

Fig. 2: The plus/minus (X...) of NHL players with 8 bins.

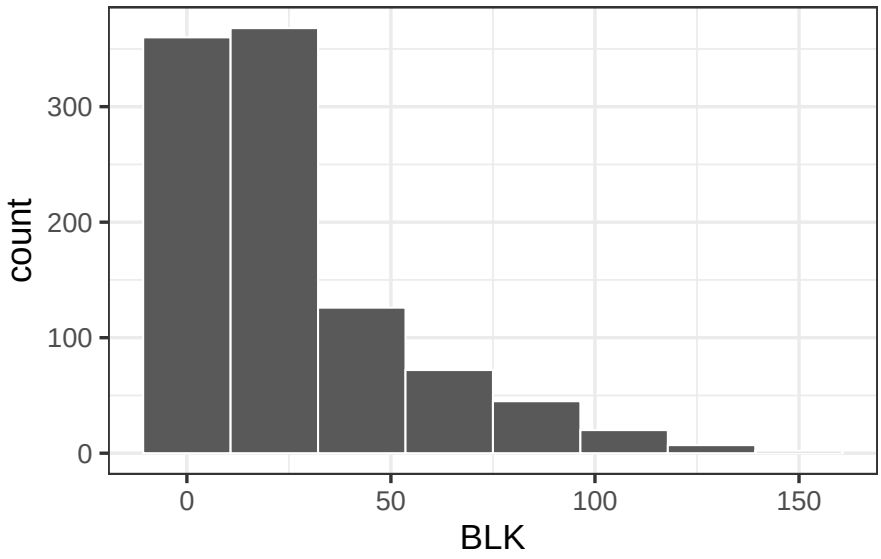

Fig. 3: Number of blocks made by NHL players with 8 bins.

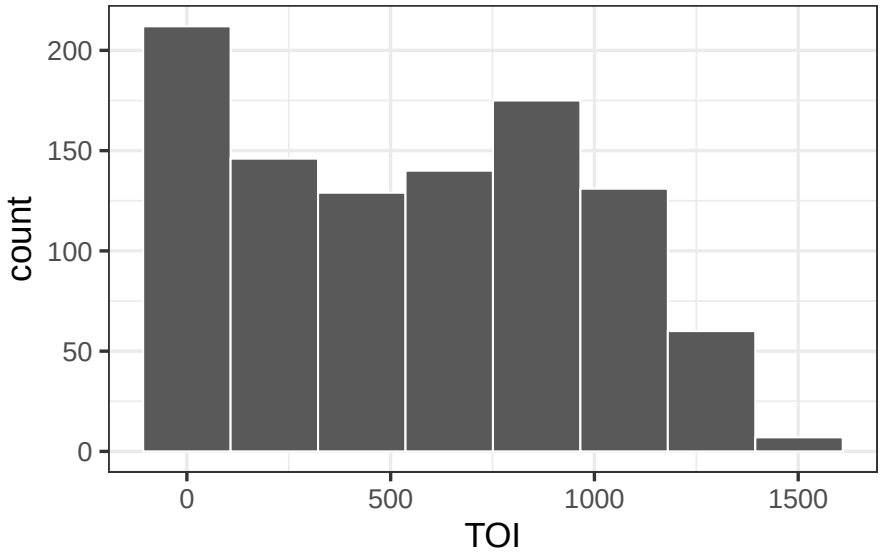

Fig. 4: Total time on ice (minutes) for NHL players with 8 bins.

The skewed distributions for Fig. 1 and Fig. 3 make sense as even in the NHL there are the best of the best that will have more points and blocks than the other players. These two variables also cannot be lower than 0, which contributes to this distribution shape. Furthermore, the better players also just get more ice time, which could improve their other statistics. From Fig. 4, we can see that ice time is also skewed. Therefore, it is reasonable that Fig. 1 and Fig. 3 are skewed. However, plus/minus can be positive or negative. Given this, Fig. 2 displays a normal distribution about 0 with long tails due to abnormally good or bad players.

Dividing each variable by age group displays similar distributions as Fig. 5, Fig. 6, and Fig. 7 all display below.

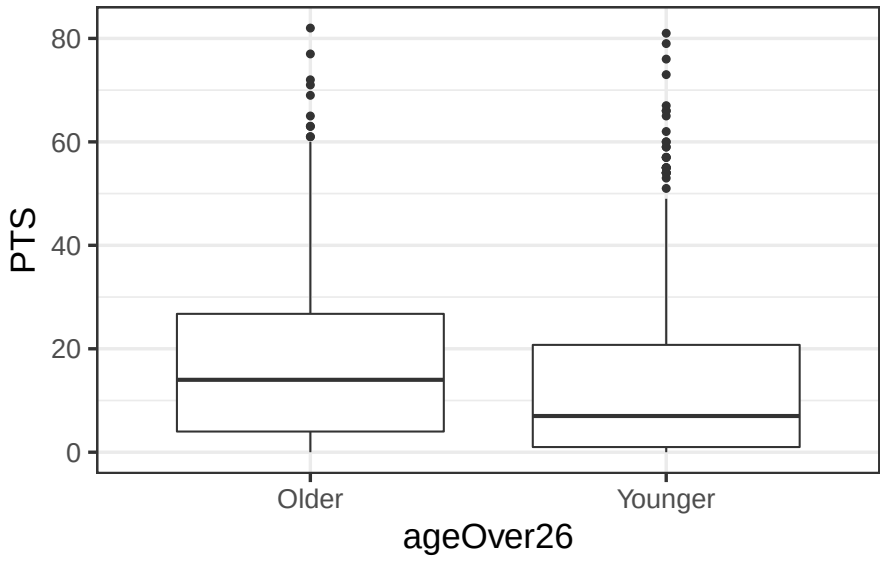

Fig. 5: Distribution of PTS against age group

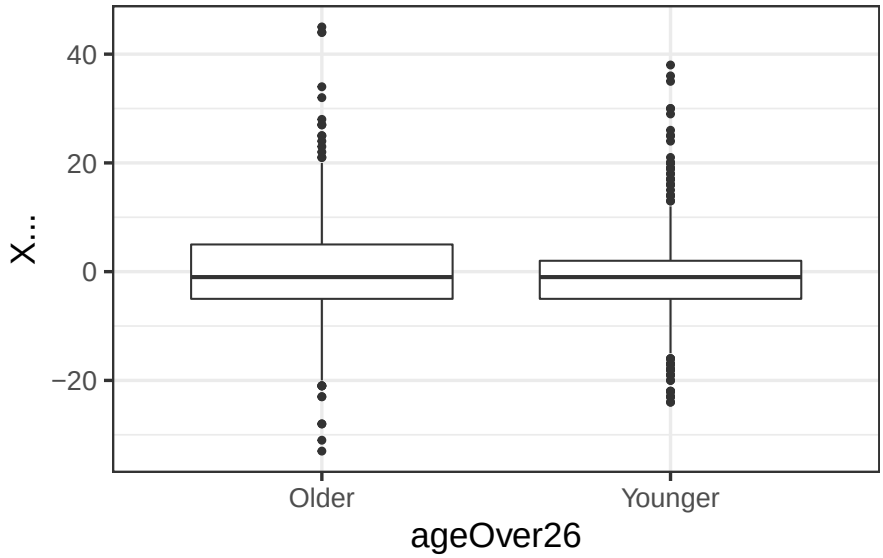

Fig. 6: Distribution of plus/minus (X...) against age group

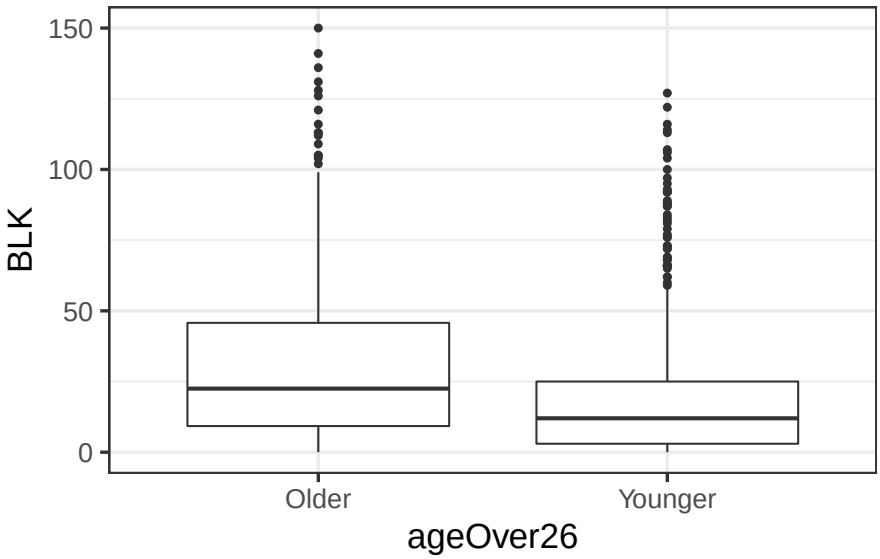

Fig. 7: Distribution of BLK against age group

The distribution of points is on average higher in the older age group than the younger age group, with averages of 17.648 and 13.709 points respectively. However, the standard deviation of points in each age group is about the same at 16.295 and 16.259 for older and younger players respectively.

Similar to points, the plus/minus (X...) of the older age group is higher at 0.042 than the younger age group with -0.773. A standard deviation of 10.199 in the older age group is higher than 8.611 for the younger players, displaying a smaller variability in the plus/minus of younger players.

Lastly, the average blocks for the older players is 31.872 with a deviation of 29.509, while younger players average 20.101 blocks with a standard deviation of 23.784.

Overall, despite a similar range between the two age groups, there are numerical differences between the performance of younger and older players. Specifically from a preliminary analysis, older players on average seem to perform better, despite less variability in younger players.

With an generalized average performance statistics μ, the original null and alternative hypotheses were

$$H_0 : \mu_y = \mu_o$$

and

$$H_A : \mu_y > \mu_o$$

Since the sample size satisfies the bootstrap assumptions, a bootstrap interval for the difference between age groups for each player statistic can be constructed.

A bootstrap test for points presents a 95% confidence interval of the difference in mean points between the two age groups to be from 1.894 to 5.959. A similar analysis of plus/minus (X...) and blocks produce respective intervals of −0.365 to 2.019 and 8.418 to 15.224. Since zero lies within the interval for plus/minus, it is possible that there is no true difference between age groups.

This suggests that older players are on average more likely to have a higher number of points and blocks when compared to the younger players, but not necessarily for plus/minus.

A linear regression analysis of each variable against age can be seen in Fig. 8, Fig. 9, and Fig. 10.

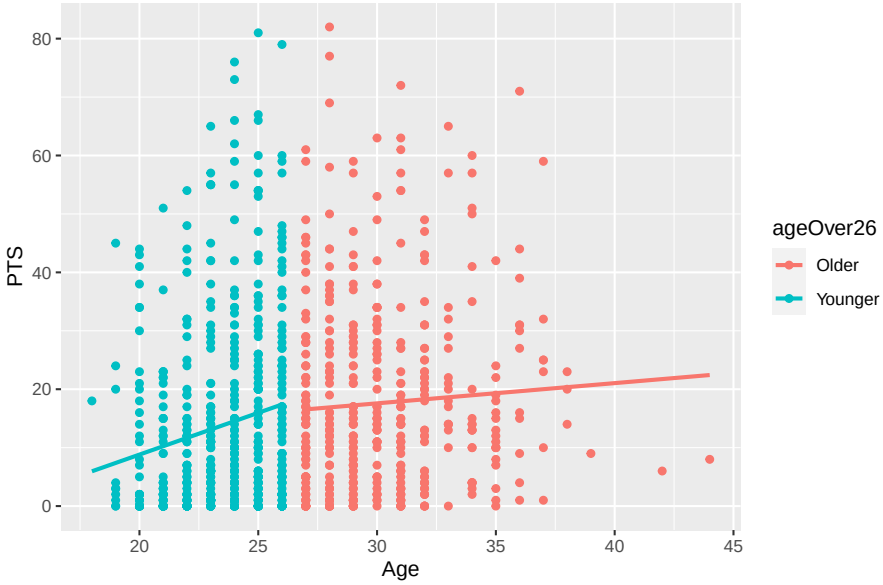

Fig. 8: Relationship between Age and Points divided by age group

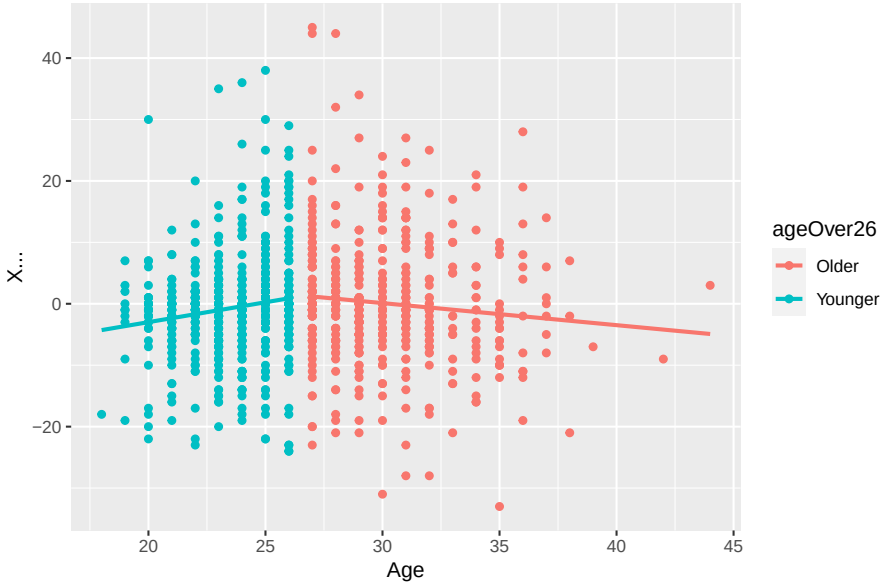

Fig. 9: Relationship between Age and X... divided by age group

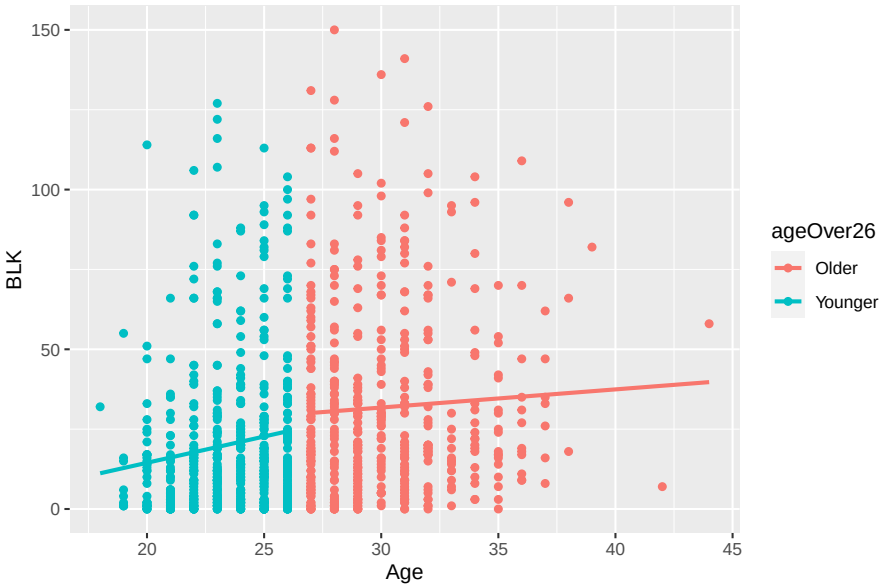

Fig. 10: Relationship between Age and Blocks divided by age group

Computing the linear regression line, Fig. 8 shows that 0.142% of the variability of points is explained by age for the older age group, while 2.64%is explained for the younger age group. Furthermore, the slope of the older age group regression line for every one year increase in age, the total point count increases by 0.3461 with an intercept of 7.199 while they are 1.429 and -19.782 for the younger age group.

For plus/minus (X...) in Fig. 9, 0.777% of the variability is explained by age for the older age group, while 1.879% of the variability is explained by age for the younger age group. Furthermore, the slope of the older age group regression line for every one unit increase in age, the players' plus/minus decreases by 0.359 units with an intercept of 10.882. For the younger age group the slope is 0.647 while the intercept is -15.938.

Finally, the regression of blocks in Fig. 10 shows that 0.077% of the variability in blocks is explained by age for the older age group, while 1.561% is explained by age for the younger age group. The slope of the older age group regression line for every year increase in age, the total block count increases by 0.568 with an intercept of 14.721. The slope of the younger age group is slightly larger at 1.644 with an intercept of -18.417.

These results seem to indicate that the linear model might not fit these relationships. To determine the truth of this, the residual plots would display some pattern, since we expect a linear model to produce random residuals about 0.

When plotting the residuals of total points in Fig. 11, there appears to be a linear spread of points intersecting the $y = 0$ line, confirming that a linear model is not applicable to this relationship. A similar pattern is seen for plus/minus in Fig. 12, suggesting again that a linear model is not applicable to this relationship. Finally,

the residuals of total blocks in Fig. 13 indicate that a linear model is not valid.

Discussion

We evaluated the hypothesized relationship between player performance and their age. Specifically, for the 2021-2022 season we looked at each players number of points, plus/minus, and number of blocks. Determining the relationship between each of these variables and age can provide useful insight on the performance of older and younger NHL players.

With a preliminary analysis on each variable, we noticed most players gravitating towards 0 in each variable, with points and blocks having righ-skewed distributions. This makes sense as a large portion of NHL players do not see much ice time, if any at all. To confirm this we checked the total time on ice for players, which also displayed a skewed right distribution. This was less skewed then the other data, which makes sense as there are a lot of in-between players who get ice time but only to rest the better players. Therefore, we would have a sizable chunk of players who get ice time but not a proportionate amount of stats. We proceeded with a linear model analysis of each variable (points, plus/minus, and blocks) against age, where we also highlighted players older and younger than the median age.

We found the linear model for points, plus/minus, and blocks accounted for 2.5%, 0.12%, and 5.0% of the variability in each model respectively. However, we found p-values of 4.6×10^{-7}, 0.27, and 7.4×10^{-13} respectively. This indicates that both points and blocks do have a relationship with age, but for all three a linear regression is not an accurate model for their relationship with age.

The residuals for each variable, seen in Fig. 11, Fig.12, and Fig. 13, agree with this conclusion. They display a linear pattern rather than the randomized pattern we should expect for a true linear relationship. However, a limitation of the study is that the limited control over the data. There were multiple duplicates due to mid-season transfers, where their statistics were recorded separately for each team. Further analysis would involve identifying and accounting for discrepancies such as these.

Even though a linear regression might not best model this relationship, the data shows that there is still a relationship. We originally predicted that younger players would bring more energy to the ice and would have better points, plus/minus, and blocks. However, our analysis show that older players tend to perform better. We believe this could be because older players only remain in the league if they are good enough to really contribute. The younger age group is most likely a mixed bag of the stars we will come to love and bench-warmers, who are not quite good enough to survive. This weeding and adapting to the league is our explanation for the relationship we uncovered but further analysis is necessary to determine the best model

for these relationships. This further analysis would be useful to NHL teams as it would help determine how much they should invest in older versus younger players.