Hockey Analytics

Aubreigh DeFriece

Time on ice plays a critical role in scoring goals. This year, Alex Ovechkin surpassed Wayne Gretzky's record and became the highest-scoring skater in the NHL. This historic event sparked my interest in comparing how time on ice and points correlate with one another. I hypothesized that players with more time on ice would have higher points per season. The game has changed over the years, as any game does. Therefore, I took players' statistics from the past ten seasons (2015-2025) to apply a linear regression to be able to predict how correlated the two factors are to each other. I plotted the different players' time on ice and points scored in the season. Overlapping that data is the regression line that showed a positive correlation between time on ice and points. It was found that more time on ice led to more points scored. However, too much time on ice decreases the number of goals scored for most skaters. This suggests that too much ice time may lead to fatigue or loss of concentration. The analysis highlights the importance of time on ice. While increased time can increase opportunity, too much can cause exhaustion and decrease performance. Teams should consider these findings when allocating time between players. To start the data processing, I created a visual to show the top 10 goal scorers in the 2024-25 season (see Table 1).

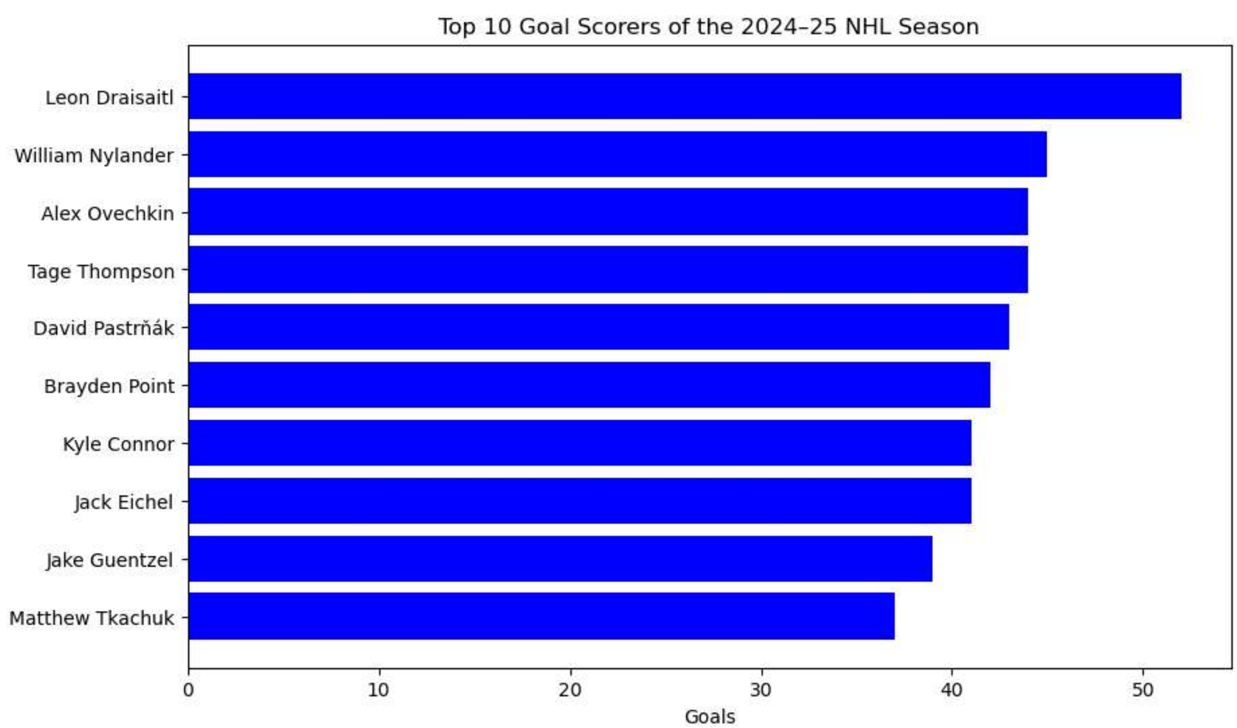

Table 1: A table of data showing the top 10 goal scorers for the 2024-25 season

By using Hockey-Reference.com, I was able to gather the skater statistics of each player in each season. The key features used were time on ice (TOI) and points (PTS). The preprocessing steps taken were cleaning the data, extracting missing values, and scaling. Players with incomplete records or who played fewer than 50 games were excluded, so outliers would not skew the data. Any missing values were extracted from the dataset to avoid errors in plotting the data. Lastly, standardization was applied to ensure each feature was contributing equally to the process. Some assumptions were linearity, independence, and normality. It was assumed that there was a linear relationship between the independent (time on ice) and dependent variable (goals scored). It was also assumed that statistics were independent of each other. Lastly, it was assumed that the residuals followed a normal distribution. I chose linear regression because of its interpretability and effectiveness in being able to model the relationship between the two variables. It provides an effective illustration of the data being compared. The regression did not require any standardization because the data was already shown on the "same level."

Once I imported the CSV files, I found that some column names had changed during the transfer process. To maintain the consistency of the data, I renamed the relevant columns to standardize the dataset. I removed any rows with missing values in the TOI and PTS columns to ensure accuracy. This step was necessary to prevent errors. The TOI data was initially in a 'minutes: seconds' format. I converted this into a decimal format representing total minutes. This makes comparing the numbers easier. To prepare the data for the regression, I converted the PTS to numeric types. This ensures that all entries were in a suitable format for the operations.

After cleaning each season's datasets, I combined them into a single data frame. This allowed for a full analysis across multiple seasons. I made TOI the independent variable X and PTS the dependent variable y for the regression. Utilizing scikit-learn's LinearRegression, I trained the model to understand the relationship between TOI and PTS. To assess the model's performance, I calculated the R^2 score, which shows the variance in PTS explained by TOI. Finally, I created a scatter plot of the data points and showed the regression line. This provided a visualization of the correlation between TOI and PTS across each season (see Table 2).

The linear regression was used to find the correlation between players' time on ice and points scored per season over the past 10 NHL seasons. The model showed an R^2 value of 0.288. This means that approximately 28% of the variance in points scored can be explained by the time on ice. A scatter plot was used to visualize the data, with TOI on the x-axis and PTS on the y-axis. The regression line on the scatter plot showed a positive correlation between TOI and PTS. However, the plot also revealed a plateau in points scored beyond a certain threshold of TOI, suggesting diminishing returns with excessive ice time. Residual analysis was conducted to assess the model's assumptions. The residuals appeared randomly scattered around zero, supporting the assumption of linearity.

The regression showed a positive correlation between Time on Ice and Points across NHL players from the 2015–2024 seasons. The model's R^2 value showed that a significant portion of the variance in points scored can be explained by the amount of time a player spends on ice. This suggests that players with more ice time tend to score more points. However, the scatter plot shows that beyond a certain time, it leads to a decline in performance. This could be because of factors such as fatigue.

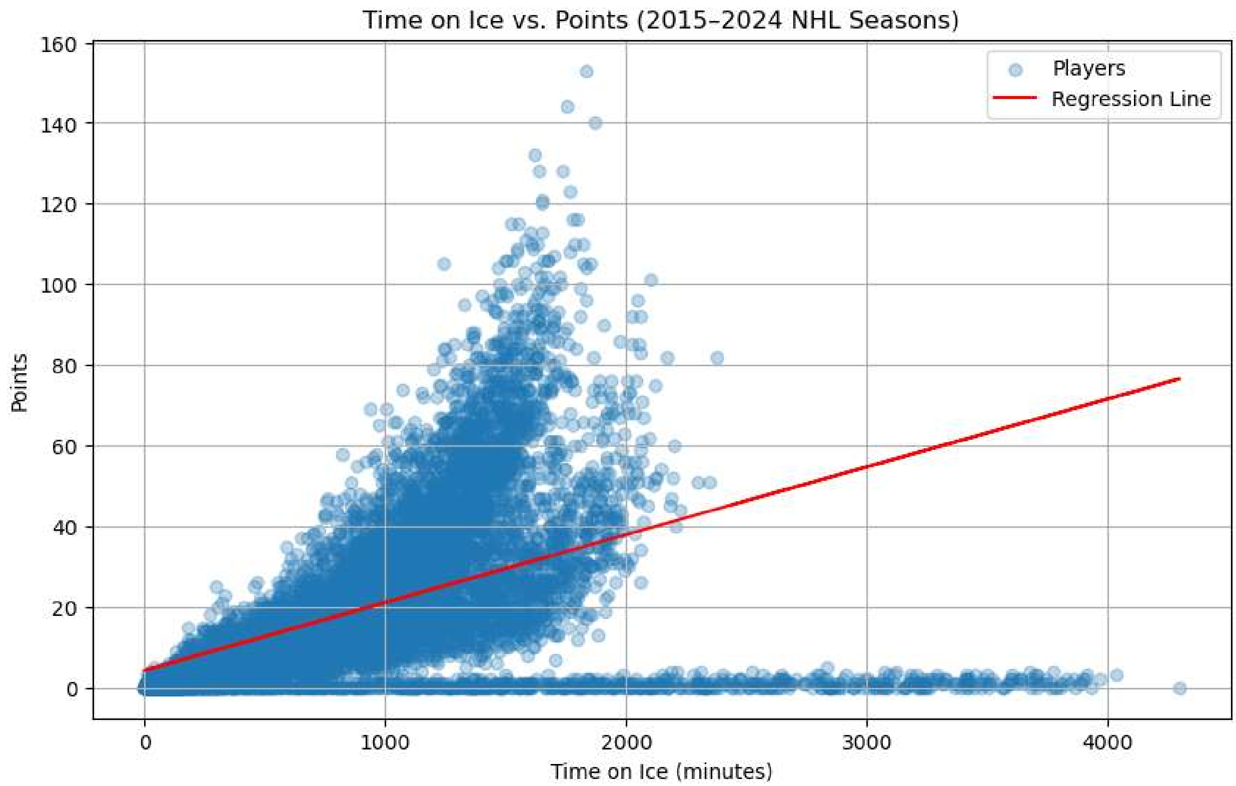

Table 2: Final linear regression of the data from 2015-2025

One of the primary challenges encountered was data preprocessing, particularly converting the TOI from a string format to a numerical value in minutes. Also, ensuring the dataset was free from missing or non-numeric values was needed for the data to be accurate. The decision to use linear regression was made because of its effectiveness in showing the relationships between two continuous variables. This project opens ideas for more complex analyses. Adding additional variables such as player position and age could provide a deeper understanding of the factors. Exploring other models or machine learning techniques might also capture the little factors observed in the data, especially in the decline in points with excessive ice time. Separating the data to analyze specific player roles or comparing forwards and defensemen separately could show more insights as well. Forwards are more likely to see an increase in points with more ice time. Defensemen may not have the most ability to score as much as the forwards do. These analyses would benefit coaches and team management in player deployment and increasing performance.

Citations

NHL stats, history, scores, standings, playoffs, Schedule & Records. Hockey. (n.d.). https://www.hockey-reference.com/

OpenAI. (2023). *ChatGPT* (Mar 14 version) [Large language model]. https://chat.openai.com/chat