Do Shots on Goal Predict Goals Scored in NHL Games?*

A Simple Linear Regression Analysis of the 2021–22 NHL Season

Hemangi Vij

October 29, 2025

This paper investigates whether the number of shots on goal predicts the number of goals scored in National Hockey League (NHL) games during the 2021–22 season. Since shots on goal represent direct scoring opportunities, understanding their relationship to goals scored is fundamental to hockey analytics. Using a simple linear regression model with shots as the predictor and goals as the response, we find that each additional shot on goal is associated with approximately 0.056 more goals. While the relationship is statistically significant ($p < 0.001$), the model explains only about 5% of the variation in goals scored ($R^2 = 0.049$), indicating that shot volume is a modest predictor of scoring outcomes. These findings highlight the role of other factors such as shot quality, special teams, and goaltender performance in determining offensive success.

1 Introduction

In ice hockey, a common assumption is that more shots on goal lead to more goals scored. A shot on goal is any attempt that would enter the net if not stopped by the goaltender either resulting in a save or a goal. Because shots on goal directly represent offensive opportunities, teams that record more shots are generally perceived as generating more scoring chances and maintaining greater offensive pressure.

However, not all shots are equally likely to result in goals. Factors such as shot quality, player positioning, defensive coverage, goal tending ability, and game situations like power plays or even strength, all influence scoring outcomes. Thus, while shot totals are simple to track and commonly cited in hockey analytics, it remains unclear how strongly they actually predict goals scored.

*Project repository available at: https://github.com/hemangivij/nhl-regression-project.

We used play-by-play data from the National Hockey League (2022) to test whether the number of shots on goal significantly predicts the number of goals scored at the team-game level. By aggregating shot level data and applying a simple linear regression model, we quantify the average effect of additional shots on scoring and evaluate how much of goal variation can be explained by shot volume alone.

The remainder of this paper is structured as follows: Section 2 describes the dataset, Section 3 presents the regression model, Section 4 reports the results, and Section 5 discusses implications and limitations.

2 Data

There are 2802 team-game observations. Each observation represents a team-game, meaning one team's performance in a single game. The sample correlation between shots and goals is 0.221.

This analysis uses play-by-play data from the National Hockey League (2022). This season was chosen because it was the most recent completed, non-lockout season available at the time of analysis, ensuring full participation of all 32 teams under consistent rules and formats. Each regular season and playoff game was included, providing comprehensive coverage of league-wide performance.

Each observation in the dataset represents a team-game performance. In total, there are 2,802 observations (one for each team in each regular-season game). For every game, events such as shots, goals, and blocked shots were recorded at the play level, then aggregated to create team-level summaries.

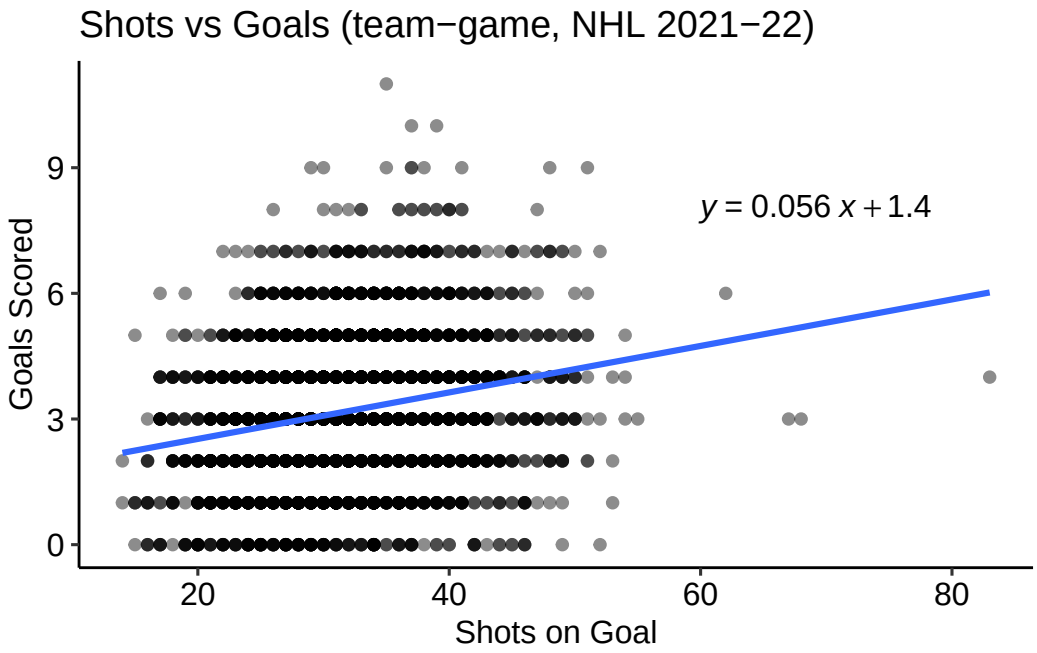

Figure 1: Scatter of shots on goal (x) vs goals (y) at the team-game level with fitted regression line.

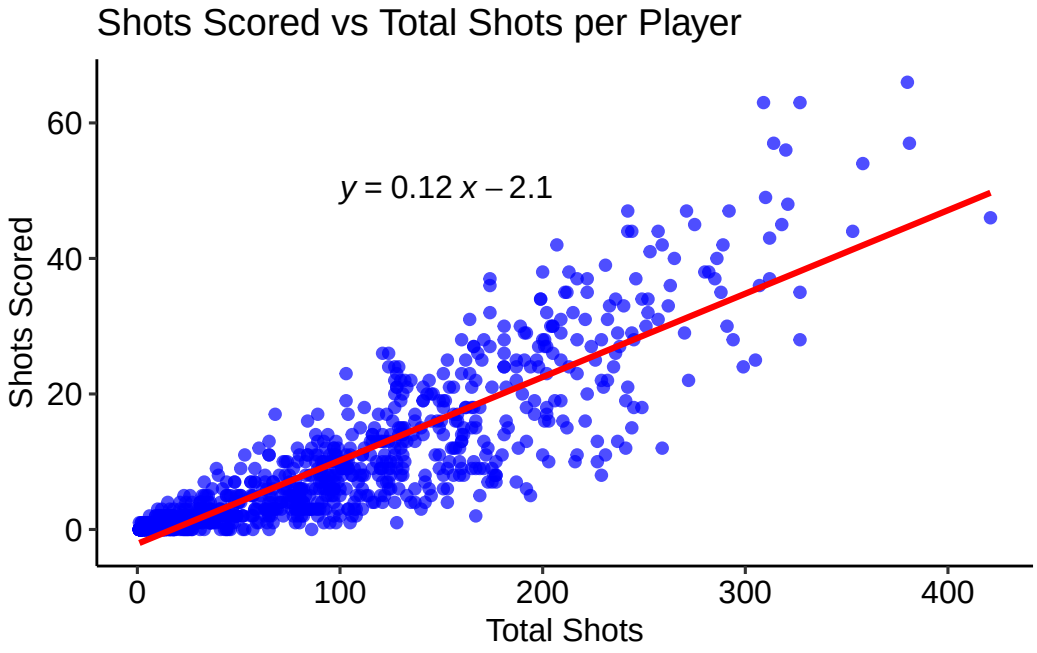

Figure 2: Shots Scored vs Total Shots per Player (NHL 2021–22) with fitted regression line.

The raw dataset records every on ice event during games such as shots, goals, penalties, and saves along with timestamps and player/team identifiers. For this study, the data was aggregated to the team-game level: each row represents one team's performance in one game. The two primary variables used were:

- **Shots on goal**: all attempts that would have entered the net if not stopped by the goaltender (including shots that resulted in goals).
- **Goals**: all successful scoring events recorded by the NHL's official scorers.

In total, the dataset contains 2,802 team-game observations, corresponding to 1,401 games across the 2021–22 season. This aggregation allows us to study how a team's shooting volume in a given game relates to its goal production in that same game.

Basic descriptive analysis shows a modest positive association between shots on goal and goals scored (sample correlation = 0.221). Figure 1 visualizes this relationship, showing that teams with more shots tend to score slightly more goals, although the relationship is far from perfect. Figure 2, which displays player-level shooting totals, was originally intended to illustrate individual variability but does not directly inform the team-game analysis and is therefore omitted in the revised version.

Limitations: Despite the richness of NHL play-by-play data, several limitations exist. Shot counts do not capture shot quality (e.g., distance, angle, or traffic in front of the net), nor do they account for special-team situations such as power plays or penalty kills. Additionally, while official NHL statistics are carefully recorded, minor inconsistencies or recording errors may occur across venues. These limitations suggest that shot volume alone cannot fully explain scoring outcomes, motivating further models that incorporate situational or player-level information.

3 Methods

To evaluate whether the number of shots on goal predicts the number of goals scored, we fit a simple linear regression model of the form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where (Y_i) represents the number of goals scored by team (i) in a game, and (X_i) represents the number of shots on goal taken by that team. The parameter (β_0) is the intercept, representing the expected number of goals when a team takes zero shots, while (β_1) is the slope, representing the expected change in goals for each additional shot on goal. The term (ε_i) is a random error term that captures unexplained variation in scoring not accounted for by shot volume.

This model was chosen because it directly reflects the research question: whether more shots on goal lead to more goals scored. The predictor (X_i) (shots on goal) is continuous and directly related to offensive output, making it suitable for linear regression. Since we are examining team-game data, each observation corresponds to a single team's performance in a single game.

Model parameters were estimated using ordinary least squares (OLS) in the **R** programming language R Core Team (2025) via the `lm()` function. Data processing and visualization were completed using the **tidyverse** Wickham et al. (2019), **ggplot2** Wickham (2016), **ggpubr** Kassambara (2023), and **patchwork** Pedersen (2024) packages.

3.0.1 Assumptions of the Linear Regression Model

Several assumptions were evaluated to ensure the validity of the regression model. It is assumed that the relationship between shots on goal and goals scored is linear, meaning that each additional shot contributes a constant expected change in goals. It is also assumed that the errors (ε_i) are independent across games, so that one team's scoring outcome does not influence another. In addition, homoscedasticity, or constant variance of the errors across all levels of shots on goal, expressed mathematically as ($\text{Var}(\varepsilon_i) = \sigma^2$) is also assumed. This condition ensures that the variability in goal scoring remains consistent regardless of how many shots a team takes. Finally, it is assumed that the errors follow an approximately normal distribution with mean zero, ($\varepsilon_i \sim N(0, \sigma^2)$), which allows for valid (t)-tests and confidence intervals. These assumptions were checked using residual plots and found that the residuals were roughly centered around zero, with a slight increase in variance at higher fitted values. This mild heteroskedasticity indicates that while the model captures the main trend between shots and goals, the precision of the standard errors may be slightly affected.

3.0.2 Model Validation

Model validity was assessed using diagnostic plots. Residuals were examined to verify that they were centered around zero (supporting linearity) and roughly constant in spread (supporting homoscedasticity). The residual plot (Figure 3) revealed mild heteroskedasticity, where residual variance increases slightly with fitted values. While this does not bias coefficient estimates, it may affect the accuracy of standard errors and hypothesis tests.

3.0.3 Limitations and Extensions

A key limitation of this model is that it assumes shots on goal are the only predictor of scoring, which simplifies the complex nature of hockey outcomes. Important factors such as shot quality, player skill, game context, and goaltender performance are not included in the analysis. Future

models could incorporate additional explanatory variables such as expected goals (xG), power-play opportunities, or goalie save percentage to capture more of the variability in goal scoring. Count-based models such as Poisson or negative binomial regression could also better represent the distribution of goals, which are nonnegative integer counts.

4 Results

We estimated a simple linear regression model of goals scored on shots on goal. The estimated intercept is ($\widehat{\beta}_0 = 1.414$), and the estimated slope is ($\widehat{\beta}_1 = 0.056$), meaning that, on average, each additional shot on goal is associated with about 0.056 more goals scored. This implies that a team taking ten additional shots in a game would be expected to score roughly half a goal more, holding other factors constant. The model explains about ($R^2 = 0.049$) of the variation in goals, suggesting that shot volume alone accounts for only a small part of goal scoring differences across games.

We used a two-sided (t)-test to evaluate the null hypothesis ($H_0 : \beta_1 = 0$) against the alternative ($H_A : \beta_1 \neq 0$). The slope estimate is positive and highly significant (($p < 0.001$)), so I rejected (H_0) and concluded that shots on goal are positively associated with goals scored. This result aligns with the intuitive idea that more offensive activity tends to create more scoring opportunities, but the small (R^2) of t value highlights that shot quantity is only one part of the story.

To check whether the assumptions of the regression model were satisfied, I examined the residuals plot (Figure 3). The residuals were roughly centered around zero, which supports the assumption of linearity, and they showed no systematic pattern, suggesting independence of errors. However, the spread of the residuals increased slightly as fitted values grew, indicating mild heteroskedasticity. In other words, the variability of prediction errors was somewhat higher for high-scoring games than for low-scoring ones. This pattern means that while the coefficient estimates ($\widehat{\beta}_0$) and ($\widehat{\beta}_1$) remain unbiased, the standard errors may be less reliable, which can make hypothesis tests or confidence intervals less precise.

Collectively, the model provides strong statistical evidence that teams generating more shots are more likely to score, but it also emphasizes the limitations of relying solely on shot volume as a predictor. The weak explanatory power suggests that other components such as player skill, player experience, special teams, or goaltending must also influence scoring outcomes. Future analyses could incorporate these variables to better capture the complexity of goal production in NHL games.

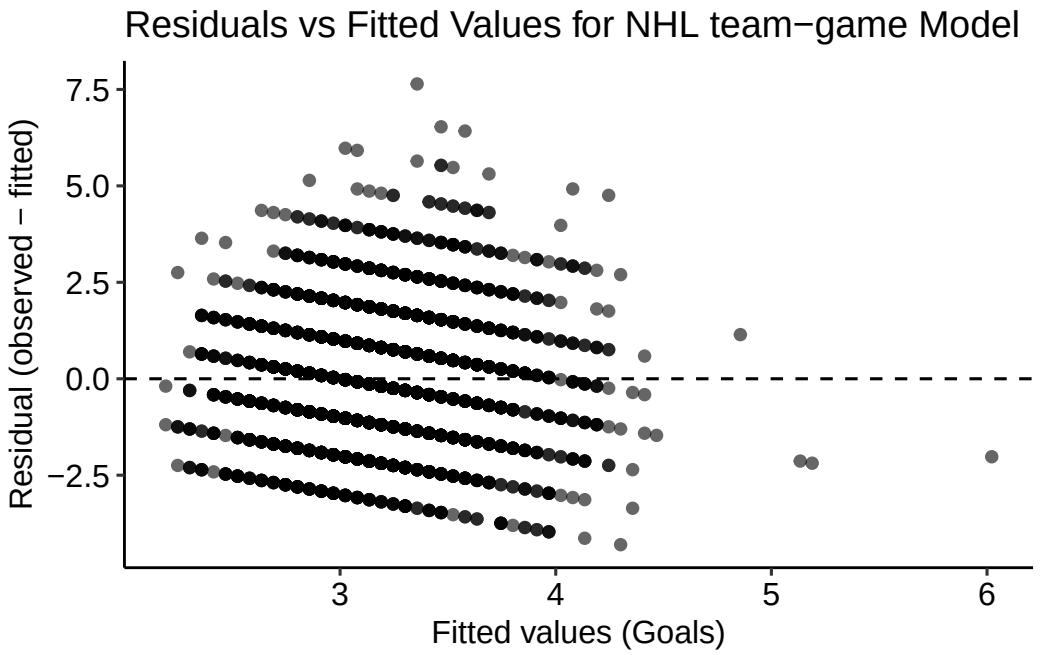

Figure 3: Residuals vs fitted values for a simple linear regression with goals as the response and shots on goal as the predictor.

5 Discussion

The presence of mild heteroskedasticity indicates that the model does not capture all of the variability in goal scoring. Although the coefficient estimates remain unbiased, heteroskedasticity can reduce the reliability of standard errors, which in turn makes statistical tests and confidence intervals less precise. As a result, while the positive relationship between shots on goal and goals scored is statistically significant, the exact precision of the estimates should be interpreted with caution. The residual plot supports the conclusion that residuals are roughly centered around zero, suggesting linearity, but their spread increases for games with higher fitted goal values, showing that prediction errors are larger for high-scoring games than for low-scoring ones.

Overall, the results support the intuitive idea that taking more shots increases the likelihood of scoring, but the relatively low (R^2) value shows that shot volume explains only a small portion of goal variation. Other factors such as shot quality, player positioning, power-play opportunities, and goaltender performance likely play major roles in determining scoring outcomes. These findings emphasize that a simple linear model can capture the general trend but cannot fully represent the complexity of offensive performance in hockey.

This study has several strengths. It uses official NHL play-by-play data from the entire 2021–

22 season, providing a large and comprehensive dataset that represents all 32 teams. The use of a simple linear regression framework also makes the relationship between shots and goals easy to interpret and communicate to a general audience.

However, there are important limitations. The model does not account for shot quality, shot type, or player context, which can significantly influence scoring outcomes. It also ignores special-team situations such as power plays and penalty kills, as well as goaltender performance. In addition, the residual analysis revealed mild heteroskedasticity, indicating that the variability of prediction errors changes across games. Although this issue does not bias coefficient estimates, it can affect the precision of statistical inference.

Future research could strengthen this analysis by incorporating additional predictors, such as expected goals (xG), power-play opportunities, or goalie save percentage, to improve model fit. More advanced count-based models, such as Poisson or negative binomial regression, may also better reflect the discrete nature of goal counts. Together, these extensions would provide a more accurate and comprehensive understanding of what drives offensive success in the NHL.

References

Kassambara, Alboukadel. 2023. *Ggpubr: 'Ggplot2' Based Publication Ready Plots*. https://rpkgs.datanovia.com/ggpubr/.

National Hockey League. 2022. "NHL Play-by-Play Data, 2021–22 Season." https://www.nhl.com/stats/.

Pedersen, Thomas Lin. 2024. *Patchwork: The Composer of Plots*. https://patchwork.data-imaginist.com/.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.