Graphics Project Final Report

36-315: Statistical Graphics and Visualization, Spring 2024
Grant Allvin, Melos Bekaj, Lay Len Ching, Silas Wang

Introduction

National Hockey League Shots (https://data.scorenetwork.org/hockey/nhl-shots.html) is a dataset that comes from the 2021-2022 National Hockey League (NHL) season. This dataset has a wide variety of quantitative and qualitative variables, with many potential areas of inquiry. The dataset contains 160,573 rows, each representing one shot made by a player, and 22 columns that each give a data point on that shot. The 22 columns provide data on the following metrics: game_ID, description of the shot, shot outcome, period of game, seconds remaining in the period, seconds remaining in the game, home team score after the shot, away team score after the shot, home team name, away team name, team that made the shot, name of goalie (if in net), whether the net was empty or not, name of primary event causer, role of primary event causer, name of secondary event causer, role of secondary event causer, how strong the game was, x-coordinate of event (in feet), y-coordinate of event (in feet), distance from shot to net, and the angle from shot to net. From this data set, variables of interest include whether or not the goalie was in the net and shot metrics such as angle and distance. We are interested in this dataset's potential research questions, such as about shot angle affecting the success rate and where on a rink players should shoot from to ensure highest probability of a goal.

We had three research questions that we wanted to explore with this dataset. First, we aim to discover any areas on the ice where goals are more likely to be scored. Therefore, we analyzed the variables of shot location, shot angle, and shot outcome. By understanding areas where shots are more likely to be successful, our results may be used to help hockey teams optimize performance. Therefore, we proposed our first research question: How do NHL teams differ in their ability to convert shots into goals, and how does this variation in goal conversion rates correlate with shot accuracy and effectiveness across different game periods?

For our second research question, we aim to analyze team performance, particularly in the context of shot outcomes. Understanding which teams excel in converting shots into goals, as well as their proficiency in generating shot opportunities, provides valuable insights into offensive strategies and player effectiveness. By identifying teams with high goal conversion rates and analyzing their shot accuracy and effectiveness, coaches and managers can make informed decisions about lineup compositions, tactical adjustments, and player development strategies. Moreover, for the NHL, such analysis contributes to enhancing the competitiveness and entertainment value of the league, as teams strive to optimize performance and deliver exciting gameplay for fans. Overall, the team performance analysis serves as a fundamental tool for driving strategic decision-making and fostering continuous improvement within the NHL. Therefore, we proposed our second research question: How do NHL teams differ in their ability to convert shots into goals, and how does this variation in goal conversion rates correlate with shot accuracy and effectiveness across different game periods?

Finally, we aim to investigate the factor of a home court advantage and whether home or away teams are favored in shot outcomes. Understanding which teams are favored has many implications for many areas of sports, such as sports betting. It is interesting whether the mere presence of playing in your home arena may increase the odds of successful shots, or whether this effect has been exaggerated or misrepresented in the media. Therefore, we proposed our third research question: Is there a "home-rink advantage" between home and away games for NHL teams, and to what extent, if any, does this effect exist?

Research Question 1

Our first research question analyzes shot outcomes based on where the shot is taken from and the angle at which it is made. Through this endeavor, we may better inform ourselves, as well as coaches and managers, about best practices for players and their shots to improve training. This analysis can then serve as a tool for players new and experienced for how to take their shots, thus enhancing their performance and thus the performance and success of one's team and the NHL as a whole.

To guide the answers of our first question, we construct a series of graphs followed by a formal statistical analysis that inform us about shot outcomes by shot location and shot angle.

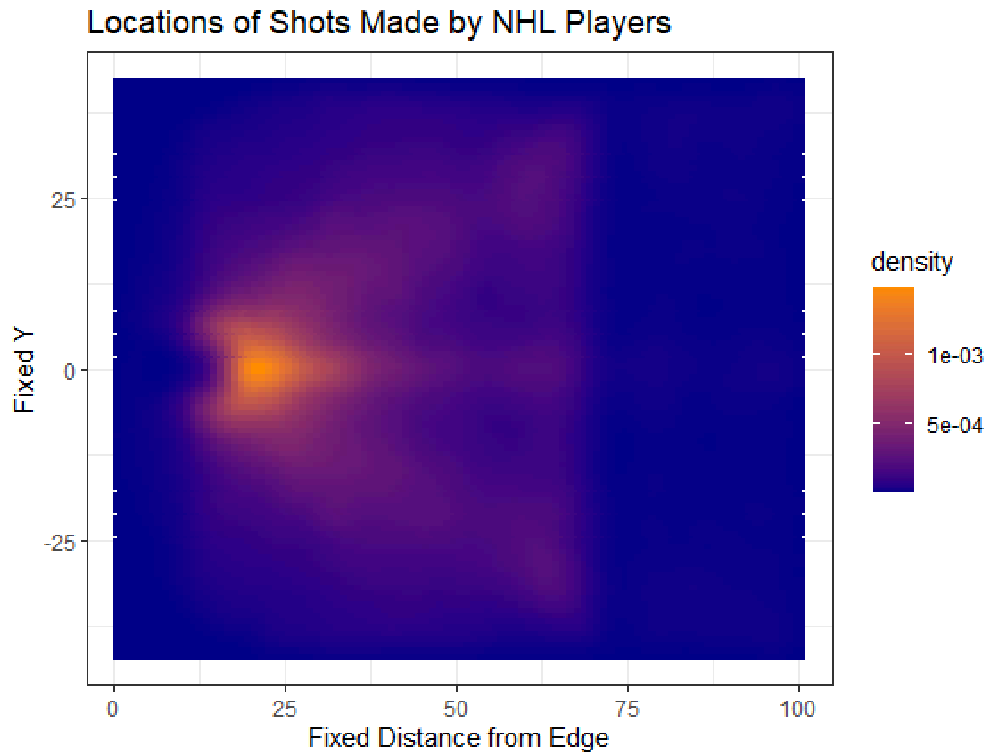
First, we construct two heatmaps that visualize the frequency of shots (and of which are goals) by the location of players on the rink. Since our questions are motivated by location-based data, we consider a heatmap to be a suitable graph.

Before this, though, we look at our data to see how location is measured, particularly the x-position x_fixed. From the data's description, the polarity of the x-position is based on which team is shooting; shots with a positive x-position are shot by the home team, while those with a negative x-position are shot by the away team. While, say, comparing how shot positions differ between home and away teams may be an interesting question to answer, it may not necessarily help to answer our questions about shot positions and angles in general, especially once we try to conduct a formal statistical analysis. With that in mind, we opt to transform this variable into a new variable x_from_edge, which measures how far a player was from either edge of the rink when making the shot with the following formula: 100-|x_fixed|. We use the 100 constant because the given rink has a fixed width of 200 units with 100 units for each team's side.

So when we make our heatmaps, we use this new transformed x-value with the original y-value.

Before we look at where goals are usually made, we deem it suitable to first look at where shots are made in general.

Plot 1: Shot Locations

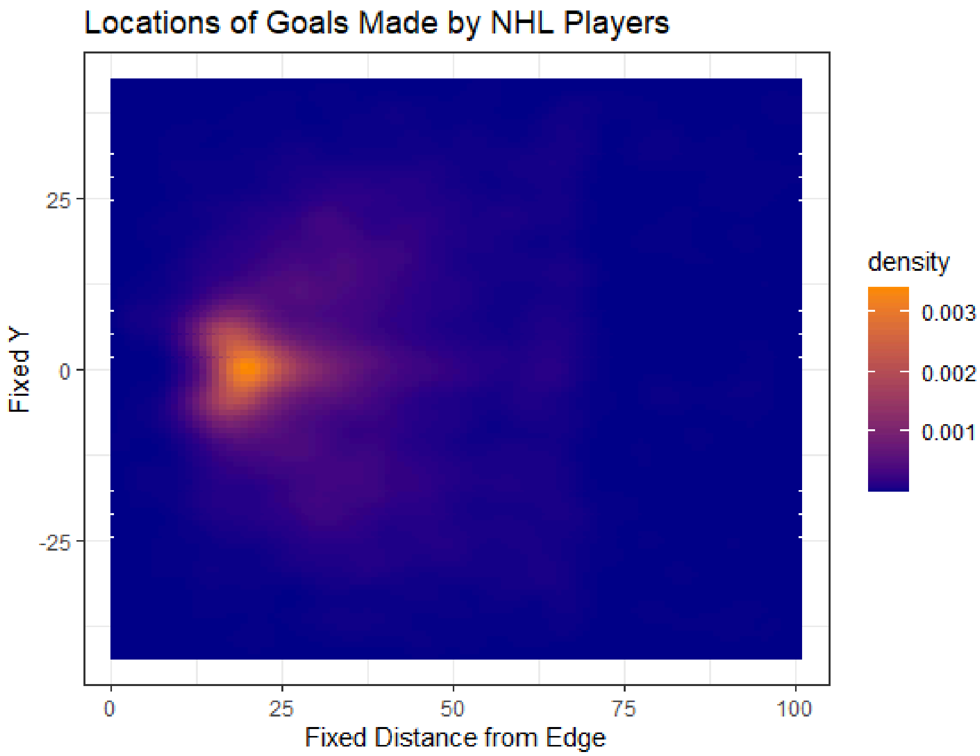


From Plot 1, it appears that most shots are made near the vertical middle of the rink (where Fixed Y is 0) and from a horizontal distance of about 20 units from the edge of the rink. Given hockey nets tend to be placed a short distance away from the edge of the rink, we may assume that this location of high shot density is around where the goal is. Going further away from the edge, density evidently decreases, but there appear to be relatively higher densities of shot locations going out from this peak in straight lines either at 0-degree or about a 30-degree angle at either side vertically. This leads us to infer that shots also tend to be made either directly in front of the net or at such a 30-degree angle (or whatever angle is made from an actual location rather than the fixed locations given in our data). This density being fairly consistent is very intriguing and perhaps is explained by a best practice for NHL players to shoot at these locations. Once the distance from the edge exceeds about 70 units, it appears as if very few shots are made, if they are made at all. It is a fairly abrupt fall-off as well, not even continuing the straight-line pattern. This may also be a result of an established best practice where making further shots is not worth the effort.

All in all, what we can infer from this graph is that hockey players tend to make shots close to the net or at some moderate distance away from such either straight through or at a 30-degree angle.

Once we have a better sense of where shots are made in general, we may then observe patterns in locations from which goals are made.

Plot 2: Goal Locations

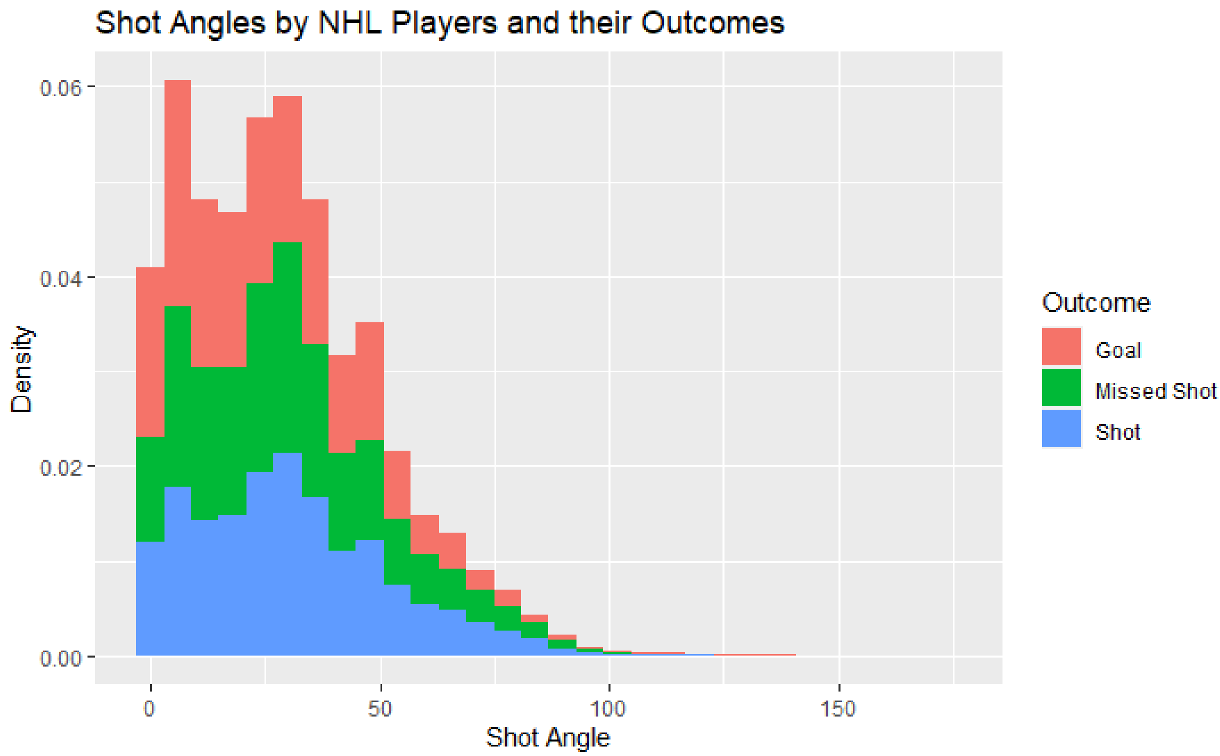


Plot 2 appears to exhibit patterns very similar to Plot 1, but with less intensity. The peak at the vertical center and about 20 units away from the edge still appears. One can get a sense of the offshoot from that peak that we see in Plot 1 as well, but it's much weaker and seems to fall-off at a shorter distance of about 40 units as opposed to 70.

With that in mind, we may infer that, in comparison to all shots made, goals tend to be made closer to what we assume to be the net and far less often at further distances.

The final part of our question asks how shot outcomes are related to the angle at which a shot is made. To make inferences that aid us in answering this question, we make a stacked histogram of shot angles with colored proportions corresponding to shot outcome. That way, we can get a good sense of the marginal distribution of shot angles and the conditional distribution of shot outcomes given said angles.

Plot 3: Shot Angles and Outcomes



By Plot 3, it appears that most shots are made at angles less than 50 degrees. Among the bars which correspond to each angle interval, each outcome appears to be about equally likely to happen, or at least not considerably more or less likely to happen than other outcomes. When conditioning on when the shot results in a goal, it appears that most of them are made at angles less than 50 degrees. This may then imply that goals tend to be made more at straighter angles, but that may be because shots in general tend to be made more often within the same angle interval.

Analysis: Logistic Regression on Scoring Goals by Shot Location and Angle

With some information from our graphs for context, we now decide to conduct a regression analysis where we develop a binary logistic regression model to predict shot outcomes. Since scoring a goal is the only outcome which scores points (which is the main objective of the game), we opt to make this model to predict whether a shot is a goal or not. Since we are interested in determining how shot outcomes are related to shot location and angle, we regress this goal indicator on the x position (our case being x_from_edge) and y position of the player, as well as the angle of the shot. While it may be better in general to regress on an x position which corresponds to that of the net, we are not certain what the x position of the net is. Since shots don't appear to be made very often behind what we assume to be the net (as observed in Plots 1 and 2), using the x-position that acts as a measurement of horizontal distance from the edge may be suitable. Because shot and goal frequencies appear to mirror vertically in our heatmaps, we infer that our y-value on its own would not be a useful predictor for our model, so we opt instead to take its absolute value when regressing on it.

```
Call:
glm(formula = goal ~ x_from_edge + abs(y_fixed) + shot_angle,
    family = "binomial", data = data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.719628   0.040596  -17.73   <2e-16 ***
x_from_edge  -0.033677   0.001092  -30.83   <2e-16 ***
abs(y_fixed) -0.030293   0.001976  -15.33   <2e-16 ***
shot_angle   -0.010243   0.000822  -12.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 63970  on 122348  degrees of freedom
Residual deviance: 59827  on 122345  degrees of freedom
  (38211 observations deleted due to missingness)
AIC: 59835

Number of Fisher Scoring iterations: 6
```

From our model summary, we infer the following:

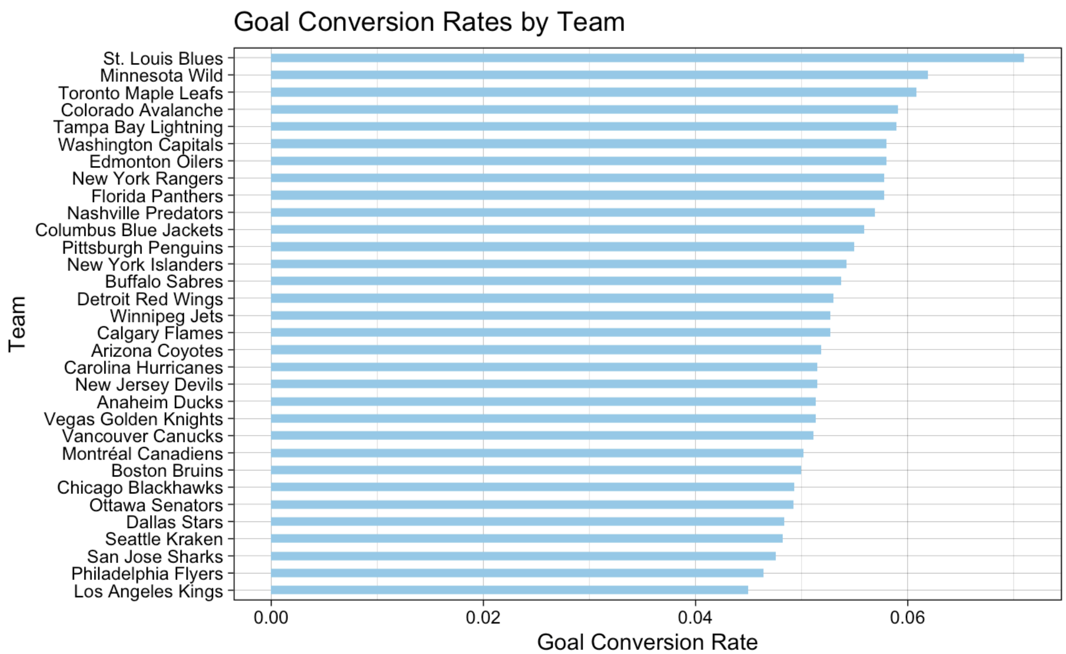
- The further away a player is from the edge of the rink horizontally (greater x_from_edge value), the less likely they are to score a goal.
- The further away a player is from the vertical center of the rink (when y_fixed=0), the less likely they are to score a goal.
- The straighter a player shoots (shot angles closer to 0 degrees), the more likely they are to score a goal.

Thus in summary, in aggregate with our visualizations and formal statistical analysis, we may infer that goals are made more often when a player is closer to the net (as close to the vertical center as possible and as close to the edge of the rink as possible - not including positions behind the net) and makes a straighter shot.

Research Question 2

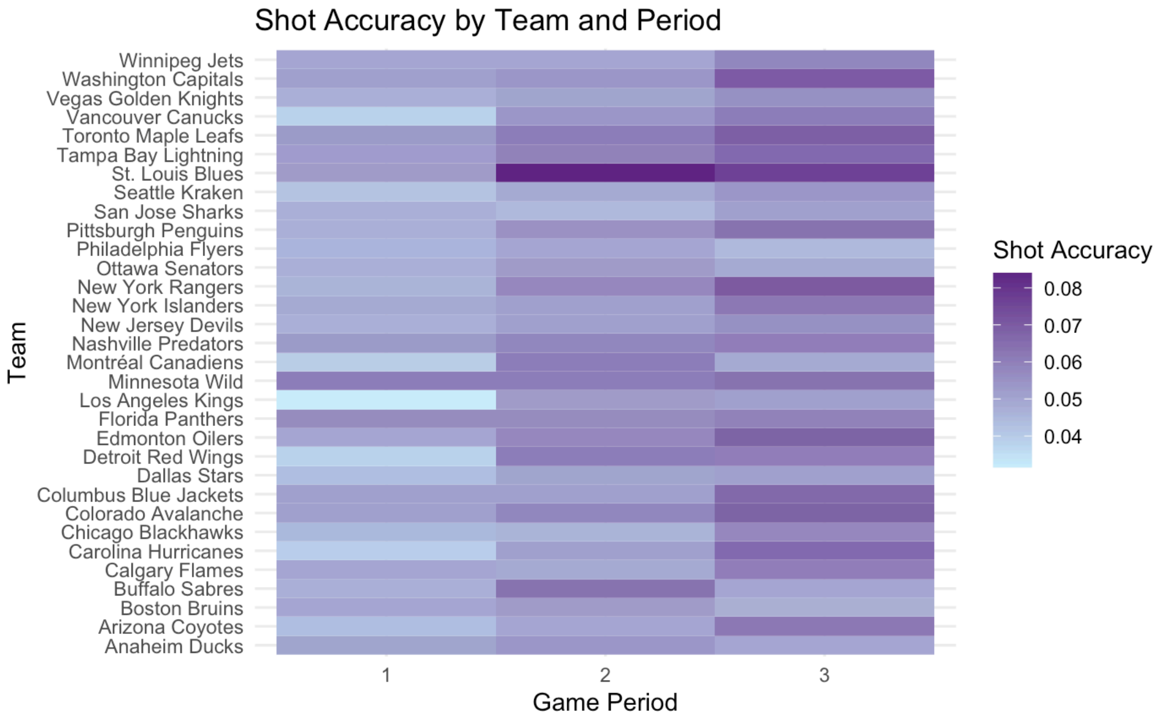
Our second research question delves into shot outcomes as an analysis of team performances. Through this endeavor, we are able to aid in the decision-making processes of coaches and managers, particularly by motivating lineup compositions, play options, and training strategies. In understanding shot outcomes by team, we can explore goal proficiency and provide valuable insights into the effectiveness of not only individual players, but offensive and defensive strategies as well. As a whole, team performance analysis can serve as a tool for strategic decision-making and contribute to enhancing the success of the NHL organization.

Plot 4: Goal Conversion Rates by Team



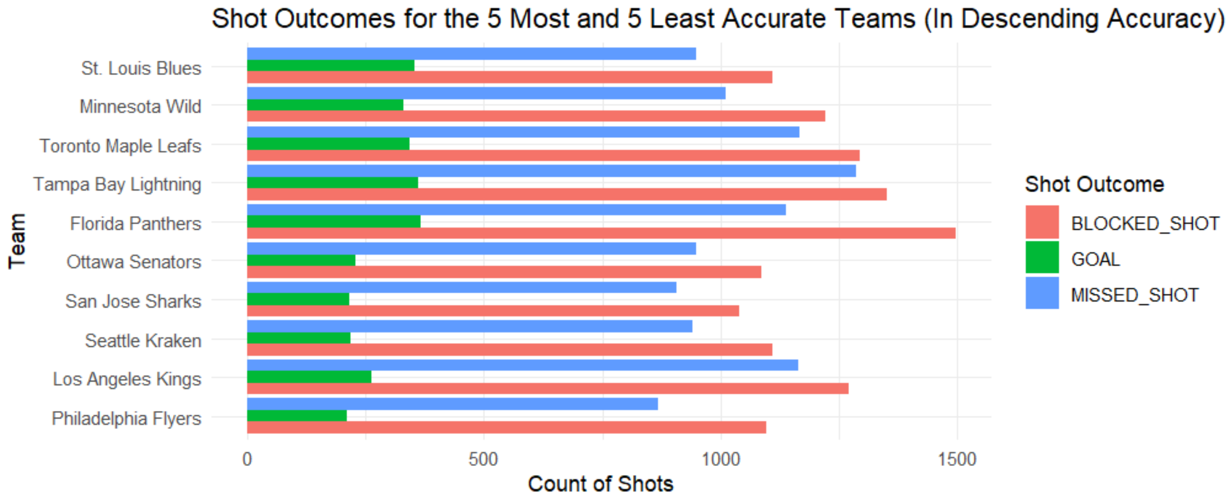
This bar plot visually represents the goal conversion rates for each NHL team, with the St. Louis Blues showing the highest conversion rate, followed by the Minnesota Wild, and the Los Angeles Kings at the bottom. This visualization connects directly to our main research question, which delves into how NHL teams differ in their ability to convert shots into goals and how this variation correlates with shot accuracy and effectiveness across different game periods. By showcasing each team's goal conversion rate, we gain crucial insights into their offensive performance and efficiency. Teams like the St. Louis Blues and the Minnesota Wild, with high conversion rates, demonstrate effectiveness in turning shot opportunities into goals, suggesting efficient offensive strategies or skilled players. Conversely, the Los Angeles Kings' lower conversion right may indicate areas for improvement in their offensive gameplay. These findings not only offer valuable insights for coaches and managers in optimizing team performance but also contribute to the league's competitiveness and entertainment value, as teams strive to enhance their offensive prowess and deliver thrilling gameplay for fans.

Plot 5: Shot Accuracy by Team and Period



This heat map visualizes team performance by shot accuracy across different game periods. It reveals intriguing insights into how teams fare in terms of accuracy throughout the game. Similar to the previous plot, the St. Louis Blues and the Minnesota Wild emerge with the highest shot accuracy overall, depicted by the dark purple color throughout the heatmap. However, a closer examination allows us to understand the nuance in trends among teams. For instance, the St. Louis Blues exhibited an exceptionally high shot accuracy in period 2, surpassing all other periods and teams. Conversely, teams like the Los Angeles Kings show variability with periods of improvement and decline in shot accuracy throughout the game. The Arizona Coyotes demonstrate a progressive increase in shot accuracy as the game advances, while others like the San Jose Sharks experience fluctuations, with period 2 showcasing their lowest accuracy. Interestingly, some teams, such as the Washington Capitals, Edmonton Oilers, Columbus Blue Jackets, and Colorado Avalanche, display stronger shot accuracy in period 3, suggesting an ability to perform more effectively as the game progresses. These findings allow us to understand the dynamic nature of team performance and fluctuations in shot accuracy, offering coaches and analysts valuable insights into the trends of a game and enabling them to adapt strategies and tactics accordingly. For instance, identifying periods of heightened accuracy for specific teams allows coaches to capitalize on these strengths, perhaps by implementing offensive plays that align with the team's peak performance periods. Conversely, recognizing periods of lower accuracy provides opportunities for targeted interventions, such as refining shooting techniques or adjusting offensive strategies to create higher-quality scoring opportunities.

Plot 6: Shot Outcomes by Team Accuracy

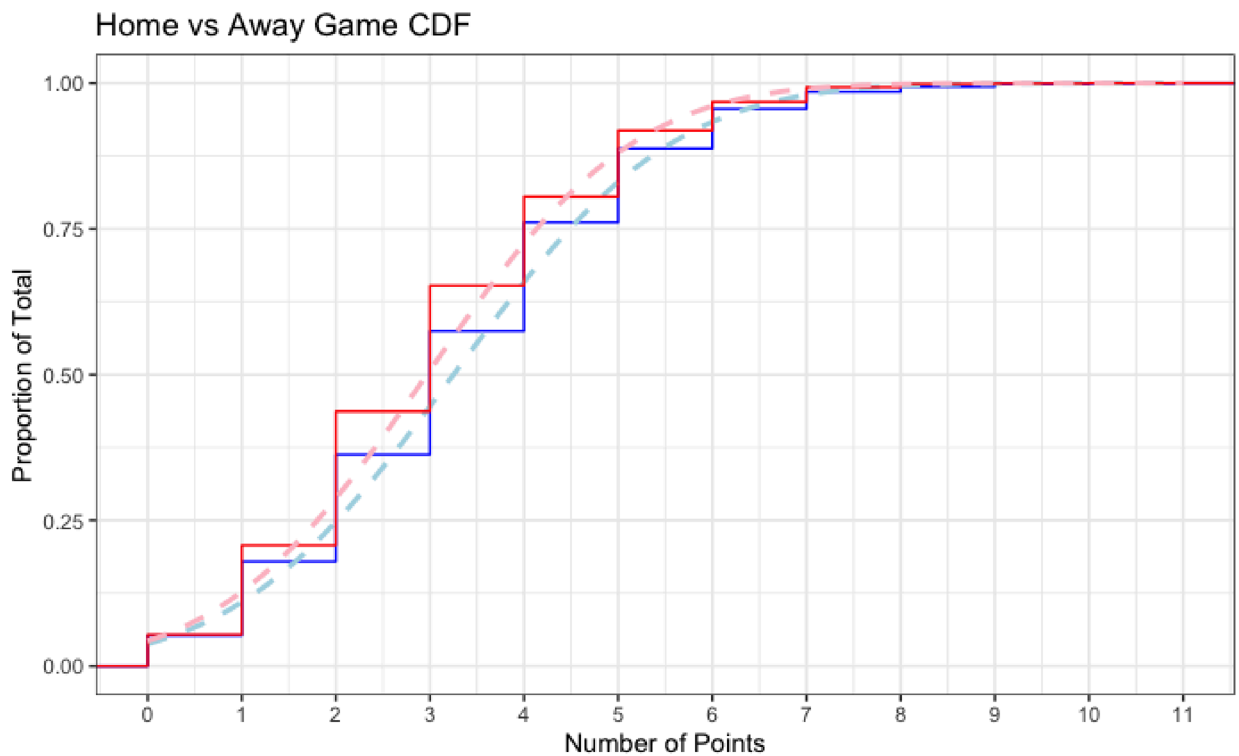


Next, this stacked barplot visualizes differences in shot outcomes by team's accuracy. We decided to focus on the most accurate and least accurate teams in terms of shot accuracy, defined as the number of a team's shots that led to goals divided by that team's total number of shots in the dataset. We identified the five most accurate teams (shown as the top 5 teams of the barplot) and the five least accurate teams (shown as the bottom 5 teams of the barplot), and they are displayed in the barplot from most to least accurate. Immediately, several trends are visible that connect to previous analysis. First, among these teams, we see strong correlations between their goal conversion scores in Plot 1 and their shot accuracy in this plot. For example, the St. Louis Blues, which lead in goal conversion, additionally lead in shot accuracy. several trends are visible. The Philadelphia Flyers have the lowest shot accuracy and the second-lowest goal conversion score, and similarly, the Los Angeles Kings, which had the lowest goal conversion score, has the second-lowest shot accuracy. Additionally, we see trends in terms of blocked shots. Among more accurate teams, we see fewer blocked shots in general and proportionally relative to their total number of shots. Overall, based on this analysis, we see areas of focus for teams wanting to improve their accuracy. Team managers may encourage players to prioritize checking for opposing team members before shooting, to overall decrease the proportion of blocked shots. Additionally, teams should work on making the shots they do take, so more practice from different areas of the ice may be useful.

Research Question 3

Our third goal for this research project is to quantify the impact of home versus away gameplay in the NHL. Understanding the influence of the home advantage can provide insights into team performance and strategies, helping teams optimize their gameplay for different environments. By examining home field advantage we can get a more holistic understanding of what factors are at play on the court. This offers a valuable perspective for fans and analysts, helping improve the depth and accuracy of discussions surrounding team dynamics, game outcomes, as well as the place fans have in their teams' success. From a business standpoint, recognizing the significance of home field advantage can influence ticket sales, marketing strategies, and fan engagement. By quantifying these impacts, the NHL can implement different strategies on and off the ice to improve player performance, competitiveness, and fan satisfaction.

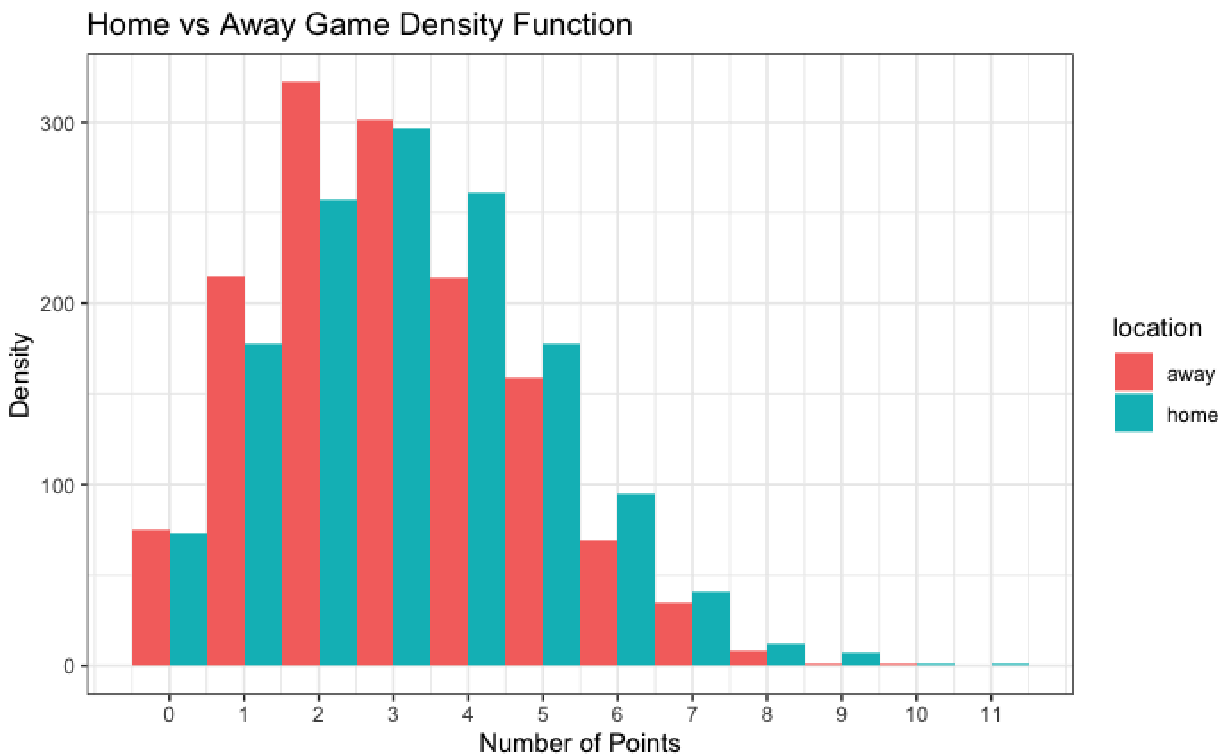
Plot 7: Home versus Away Games Cumulative Point Densities



Comparing home and away game score differences is common in sports statistics. One of the ways we chose to visualize and analyze this comparison was by using Cumulative Distribution Function (CDF) plots. These plots show the distribution of the data and can help identify differences between home (in blue) and away (in red) performances.

In this case, the means of the home and away scores were 3.25 and 2.97, respectively. This indicates that, on average, the home team tends to have a slightly higher score difference compared to the away teams, both of which followed a normal distribution. The fitted normal distribution curve for home (in light blue) lies to the right of the fitted normal distribution curve for away games (in pink), it suggests that home teams tend to have higher score differences compared to away teams.

Plot 8: Home versus Away Point Densities




```
 Asymptotic one-sample Kolmogorov-Smirnov test

data:  hockeye_trimmed$home_score
D = 0.12904, p-value < 2.2e-16
alternative hypothesis: two-sided


 Asymptotic one-sample Kolmogorov-Smirnov test

data:  hockeye_trimmed$away_score
D = 0.14933, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Using histograms allows us to visually compare the distributions of two sets of data. In our case, the distribution of home game scores is offset to the right of away games. Each histogram has a binwidth of 1 with each bin representing the total counts for that score. There are clear differences in the modes of each distribution with the mode of away games being 2 points while home games have a mode of 3 points. Home games also have higher counts for every score greater than 3 than their away game counterparts. This confirms our findings made with the CDF plot. The distribution of home game scores suggests that, on average, home teams tend to have higher scores. Using the Kolmogorov-Smirnov test, we confirm that both histograms follow a normal distribution.

Analysis: Welch Two Sample T-Test

```
  Welch Two Sample t-test

data:  hockeye_trimmed$home_score and hockeye_trimmed$away_score
t = 4.2059, df = 2790.6, p-value = 2.683e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1508792 0.4144313
sample estimates:
mean of x mean of y
 3.250535  2.967880
```

The Welch Two Sample t-test was used to determine if there is a significant difference between the means of the two groups. The test was conducted to compare the mean scores of home games and away games. The t-value of 4.2059 indicates the magnitude of the difference between the means of the two groups relative to the variability within each group. A larger t-value suggests a greater difference between the means. The p-value is 2.683 x10^-5, this very low p-value represents that there is an incredibly small probability that the differences in our data are by chance. In this case, we have strong evidence against the null hypothesis that there is no difference between the means.

The mean score for home games is 3.25, and the mean score for away games is 2.97. The 95% confidence interval for this indicates that the true difference in means is between 0.151 and 0.414. Based on the Welch Two Sample t-test, we conclude that there is strong evidence to suggest that the mean score of home games is significantly different from the mean score of away games in hockey matches. The observed difference in means is unlikely to be due to random chance alone, indicating there are real differences in the scoring patterns between home and away games. Specifically teams seem to perform better during home games, this confirms previous research on the topic.

Discussion

Overall, based on our analysis, we find several areas of focus for hockey team administrators, such as team managers, general managers, and defensive and offensive coaches. First, we find differential score potential by location on the ice. There are areas on the rink where scoring occurs more frequently, and the heatmaps demonstrate specific hotspots on the ice that may be areas of focus for players to either attack or defend. Additionally, this analysis demonstrates specific areas of high shot density, which gives team managers and officials areas of focus to study opposing teams. Next, we find trends in goal conversions, shot outcomes, and accuracy measures for different teams. We can gain some insight from examining stronger teams, and we conclude that teams wishing to improve should focus on their blocked shots, as above all, blocked shots remain the highest category of shots made by teams. Additionally, we focus on the differential trends in accuracy by period. The results point to specific periods for each team of high and low shot accuracy, identifying specific practices and plays that should be gleaned and improved. Finally, the third analysis demonstrates differential scoring by a team's home vs. away status. We identify that home teams overall have significantly higher scoring than teams that are playing away, which adds to the complexity of who would win a hockey game. This result can influence team development by encouraging teams to practice on a wide variety of rinks to ensure that proper plays and strategies may be applied to all the rinks they may play on.

In total, the dataset has numerous significant findings that may be used to optimize a hockey team's performance. While these reflect some of the findings, many insights may be gained from our analysis. Future areas of research could extend our understanding of factors influencing goal-scoring in hockey. Building on the findings regarding shot location and its impact on goal-scoring likelihood, further research could delve into optimizing shot placement on the ice. This could involve exploring specific zones or angles within the rink where shots are most effective, considering variables such as goalie positioning and defensive strategies. Another question could involve investigating the relationship between player positioning, movement patterns, and goal-scoring opportunities could provide valuable insights into offensive strategies. In analyzing how players navigate the ice to create shooting opportunities and exploit defensive weaknesses, we could inform coaching tactics and player development. Finally, we could examine goalie performance metrics in relation to shot outcomes in order to uncover insights into effective blocking strategies. We could focus on goalie positioning, reaction times, and block percentage in different game situations to identify patterns and trends that contribute to defensive success.

By exploring these topics in further details, we can deepen our understanding of the relationships between player performance, strategic decision-making, and environmental factors influencing goal-scoring in hockey. These insights have the potential to inform coaching practices, player development initiatives, and strategic decision-making within the sport, ultimately improving the competitiveness and excitement of hockey.