Ice Hockey: Sports Analysis Project

Group 1

Bevan Stanely

Nikhil Verma

Shantanu Sengupta

Abhishek Kumar

Gubbala Hari Priya

Lakshay Sharma

Sunku Vydehi

2023-11-27

Table of Contents

[1. Executive Summary 2](#_Toc1)

[2. Problem Statement 3](#_Toc2)

[2.1 What are the critical factors predicting a player’s performance? 3](#_Toc3)

[3. Proposed Solution 4](#_Toc4)

[3.1 Variable Significance 4](#_Toc5)

[3.1.1 Individual Characteristics 4](#_Toc6)

[3.1.2 Performance Characteristics 5](#_Toc7)

[3.2 Limitations 5](#_Toc8)

[4. Timeline 6](#_Toc9)

[5. Analysis 7](#_Toc10)

[5.1 Exploratory Data Analysis 7](#_Toc11)

[5.2 Regression Analysis 12](#_Toc12)

[5.2.1 Interpretation 19](#_Toc13)

[5.3 Applications of the model 21](#_Toc14)

[References 24](#_Toc15)

# 1. Executive Summary

Ice hockey is a fast-paced team sport renowned for its combination of speed, physicality, and skill, predominantly played in cold-weather nations like Canada, the USA, Russia, Sweden, and Finland. Women’s ice hockey, a variant of the sport played exclusively by female athletes, shares similarities with men’s hockey but boasts its distinct history. Since 1990, the International Ice Hockey Federation (IIHF) has organized the IIHF World Women’s Championship, showcasing the growth of women’s ice hockey.

The USA took the championship in the current year, with Canada coming in second. Since the inception of the competition, they have been regularly among the winners. The first eight championships were won by Canada, whereas the United States dominated in the more recent editions, taking home ten out of the last 14 championships([“Women’s World Championship Statistics” n.d.](#ref-WomenWorldChampionship)).

The main objective in ice hockey is to score goals by shooting a puck into the opposing team’s net. A standard game comprises three periods, each lasting 20 minutes, with intermissions for player rest and ice maintenance. Victory goes to the team with the most goals after these three periods.

In a typical lineup, each team fields six players simultaneously on the ice, featuring a goaltender (goalie) and five skaters. These skaters are typically divided into three forwards (left wing, center, and right wing) and two defensemen.

To maintain fairness and uphold the competitive spirit of the game, ice hockey employs an extensive set of rules and penalties. Players who commit infractions can be temporarily sidelined in the penalty box, leading to power plays (when one team has more players on the ice) or penalty kills (when a team is shorthanded). These regulations play a crucial role in preserving the integrity of the sport.

Our analysis involves the single-year data of the 2023 IIHF World Women’s Championship. We will address the individual player performance with linear regression and identify the key attributes affecting the performance.

|  |
| --- |
| ([“Basic Ice Hockey Rules Sportsa Articles” n.d.](#ref-BasicIceHockey)) |

# 2. **Problem Statement**

## 2.1 **What are the critical factors predicting a player’s performance?**

The challenge is to systematically identify and analyze the key factors that reliably predict a player’s performance. This includes analyzing the physical, skill-based, psychological, and strategic elements ([Nigg et al. 2020](#X9d19b3215b5ca1f18315c8df88a1bc83304c4e4)) that contribute to a player’s effectiveness in the game. There is a need for a framework that allows coaches, teams, and players to make data-driven decisions and improve the overall competitive edge in the sport.

For this study, we only looked at the official data on the physical and game statistics of 203 forwards and defenders. The rationale behind the decision to exclude goalies from our dataset stems from the disparity in their performance metrics with forwards and defenders.

In women’s ice hockey, the publicly available data is limited in terms of how important the attribute is for the players’ performance. Furthermore, there is currently no tidy dataset available. So, we are collecting and putting together information from official documents and extending it. We also want to look at the player statistics of the participants in the 2023 Women’s Ice Hockey World Championship and identify the key performance indicators for high-performing players.

# 3. **Proposed Solution**

We will compile a tidy dataset using publicly available data from the 2023 Women’s Ice Hockey World Championship. Furthermore, we will perform regression analysis to identify the dependence of the individual variables on the player’s performance. We expect to derive critical insights that could have implications for team strategizing. The insights will also help ice hockey fans understand their favorite players better.

The attributes of the data[[1]](#footnote-2) collected are as follows

|  |  |  |
| --- | --- | --- |
| No. | Attribute | Description |
|  | **Dependent variable** |  |
| 1 | Points/Game | Average of points scored across all the games played. |
|  | **Independent variables** |  |
| 2 | Position | Player-position: Forwards or Defensemen (Goalies are ignored since their performance metrics are different) |
| 3 | Goals |  |
| 4 | Assist | Goals that were assisted by focal player |
| 5 | Country | Team played under |
| 6 | Games played | Total games played by the player in the current season. |
| 7 | Height | In cms |
| 8 | Weight | In Kg |
| 9 | Penalties in minutes | Penalty duration over the championship |
| 10 | +/- Plus/minus | Player’s goal differential while they are on the ice |
| 11 | Shots on goal(SOG) | The number of attempts made by a player to score a goal that directly challenge the opposing team’s goaltender |
| 12 | Age |  |
| 13 | Shoots (left or right) | Player handedness |
| 14 | Total minutes/s played (TM) | Total player time on ice |

Our choice of points for the dependent variable covers a player’s ability to score and the team spirit assisting their teammates in scoring. As a game, hockey uses points for ranking purposes, reinforcing our choice. The remaining independent variables were picked from the player parameters available.

## 3.1 Variable Significance

We can group the things that affect how a player plays a game into different categories.

### 3.1.1 Individual Characteristics

The individual player possesses distinct characteristics such as height, weight, age, position, and player handedness (shooting). Since they could affect individual performance, they would be useful for our model.

### 3.1.2 Performance Characteristics

Resulting from the players’ participation in the championship are variables pertaining to the outcome of their participation. Goals, assists, games played, penalties in minutes, +/- differential, shots on goals, and total minutes played are performance variables. The average of total goals scored with assists is our dependent variable. The remaining variables indicate how the players’ time on the ice was spent.

Finally, the variable country groups the players with their teammates. Ice hockey being a team sport, the player’s performance is linked to their teammates. Therefore, we could use it to assess the player performance distribution within the team.

We will use the following data sources for the data.

| Sr. no. | Data |
| --- | --- |
| 1 | Quant hockey Website ([n.d.](#ref-IIHFWorldChampionship)) |
| 2 | IIHF Official Website ([n.d.](#ref-WomenWorldChampionship)) |
| 3 | Team USA ([2023](#ref-USAPlayerStatistics2023)) |
| 4 | Team Japan ([2023](#ref-JapanPlayerStatistics2023)) |
| 5 | Team Canada ([2023](#ref-CanadaPlayerStatistics2023)) |
| 6 | Team Switzerland ([2023](#ref-SwitzerlandPlayerStatistics2023)) |
| 7 | Team Czech Republic ([2023](#ref-CzechiaPlayerStatistics2023)) |
| 8 | Team France ([2023](#ref-FrancePlayerStatistics2023)) |
| 9 | Team Finland ([2023](#ref-FinlandPlayerStatistics2023)) |
| 10 | Team Germany ([2023](#ref-GermanyPlayerStatistics2023)) |
| 11 | Team Sweden ([2023](#ref-SwedenPlayerStatistics2023)) |
| 12 | Team Hungary ([2023](#ref-HungaryPlayerStatistics2023)) |

## 3.2 Limitations

We foresee two limitations of the dataset, which should be accounted for in the analysis.

1. Small sample size compared to other tournaments
2. The tournament had two parts. 1st part was a group stage and 2nd a knockout competition, giving more match time and the opportunity to score more points for qualified teams.

# 4. Timeline

The anticipated time frame for the project tasks is as follows.

|  |  |
| --- | --- |
| Task | Due Date |
| Finalize Data Collection | 5th September 2023 |
| Mid Term Review | 3rd October 2023 |
| Final Submission | 29th November 2023 |

# 5. Analysis

## 5.1 Exploratory Data Analysis

In this section, we conduct an exploratory data analysis (EDA) to understand the distribution and relationships within our dataset. Our dependent variable of interest is “Points per Game” (avg\_points), and we aim to identify key factors influencing this variable.

To get an idea of our data, we will check the data type of the variables.

df |> map\_chr(type\_sum) |>  
 as\_tibble\_row() |>  
 pivot\_longer(everything()) |>  
 count(value) |> kable()

Table 1: The count of different data types read.

| value | n |
| --- | --- |
| chr | 4 |
| dbl | 11 |
| time | 1 |

The data set includes four categorical variables, eleven numerical variables, and a unique time-varying variable. The time variable is converted into seconds. Furthermore, we will also eliminate variables related to the dependent variable to avoid overfitting the linear model to the data.

# Drop since they are related to the dependent variable  
df = df |> select(!c(games\_played\_gp,goals\_g, assists\_a, points))  
# Convert duration to seconds  
df = df |> mutate(total\_minutes\_as\_duration =  
 as.numeric(total\_minutes\_as\_duration, unit = "secs")) |>  
 # Rename some columns for easier analysis  
 rename(  
 avg\_points = points\_per\_game\_ppg,  
 penalty\_dur = penalties\_in\_minutes\_pim,  
 duration = total\_minutes\_as\_duration,  
 g\_diff = plus\_minus\_goal\_differential,  
 sog = shots\_on\_goal\_sog  
 )

Subsequently, we generate a statistical summary of all variables to examine their distribution.

Table 2: Statistical Summary of Numerical Columns

| variable | min | median | mean | stdev | q25 | q75 | max |
| --- | --- | --- | --- | --- | --- | --- | --- |
| avg\_points | 0 | 0.25 | 0.3911330 | 0.4722867 | 0.0 | 0.571 | 2 |
| height | 147 | 168.00 | 168.4482759 | 6.4201905 | 164.5 | 173.000 | 190 |
| weight | 42 | 66.00 | 65.7044335 | 6.8282559 | 62.0 | 70.000 | 85 |
| penalty\_dur | 0 | 2.00 | 2.8620690 | 4.5041993 | 0.0 | 4.000 | 35 |
| g\_diff | -10 | -1.00 | -0.0098522 | 4.3634283 | -3.0 | 2.000 | 14 |
| sog | 0 | 7.00 | 9.1477833 | 8.0054333 | 3.0 | 14.000 | 42 |
| age | 16 | 23.00 | 23.6551724 | 4.3667636 | 20.0 | 27.000 | 36 |
| duration | 0 | 5659.00 | 5982.4827586 | 9072.3116359 | 3762.0 | 7289.000 | 129946 |

plot.corr(df, "pearson")

|  |
| --- |
| Figure 1: Correlation Matrix of all the Variables |

The relationships between the variables are reflected in the correlation matrix. It is evident that there are no independent variable pairs with an absolute correlation greater than 0.7. Therefore, we can proceed with our features because of the absence of multicollinearity.

We will now explore the DV-IV relationships in greater detail. In the XY scatter plot of the dependent variable with each independent variable, we use smoothing to find the best curve. This would enable us to conduct a visual analysis of the linearity of the relationship.

df |> select(where(is.numeric)) |>  
 gather(yvar, val,-avg\_points) |>  
 ggplot(aes(val, avg\_points)) +  
 geom\_point() +  
 stat\_smooth(geom = "smooth") +  
 facet\_wrap(yvar ~ ., ncol = 2, scales = "free\_x") +  
 labs(x = "Numeric Independent Variables", y = "Average Points(DV)", title =  
 "IV vs DV Scatter Plot")

|  |
| --- |
| Figure 2: The figure shows the relationship between the each numeric independent variable with the dependent variable average points. |

All variables, except duration, have an approximately linear relationship with average points. The occurrence of an outlier in the duration results in a significant deviation from linearity. After further investigation, we conclude that the anomaly is the result of human error in data collection.

df |> select(Name, country, duration) |>  
 filter(duration == max(duration)) |>  
 kable()

Table 3: The outlier in the duration data.

| Name | country | duration |
| --- | --- | --- |
| Emma Ingold | Switzerland | 129946 |

To address the anomaly, we rectify the duration for Emma Ingold and reassess the linear correlation between the duration and the dependent variable.

df = df |> mutate(duration = replace(duration,  
 Name == "Emma Ingold",  
 as.numeric(  
 as.difftime("01:30:24",  
 format = "%H:%M:%S",  
 units = "secs")  
 )))  
df |> select(avg\_points, duration) |>  
 ggplot(aes(duration, avg\_points)) +  
 geom\_point() +  
 stat\_smooth(geom = "smooth") +  
 labs(x="Total Player time in Seconds", y="Average Points(DV)")

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

|  |
| --- |
| Figure 3: Duration vs Average Points scatter plot after replacing the outlier. |

After fixing the outlier in the duration variable, the linear relationship becomes more apparent.

Data quality and consistency are ensured by a final check of the correlation matrix and box plots.

# Plot correlation heatmap  
plot.corr(df,"pearson")

|  |
| --- |
| Figure 4: Correlation Matrix after fixing the outlier |

We can see that fixing the outlier increased the correlation between the duration and average points from 0.09 to 0.53. This increase may prove advantageous to our modeling efforts.

df |> select(-where(is.numeric), -Name, avg\_points) |>  
 gather(yvar, val,-avg\_points) |>  
 ggplot(aes(reorder(val, avg\_points), avg\_points, fill = val)) +  
 geom\_boxplot() +  
 facet\_wrap(yvar ~ ., ncol = 2, scales = "free\_x") +  
 theme(legend.position = "none") +  
 labs(x = "Categorical Independent Variables", y = "Average Points(DV)", title =  
 "IV vs DV Box Plot")

|  |
| --- |
| Figure 5: The figure illustrates the connection between the categorical independent variables and the dependent variable’s average points via box plots. Of the three IVs, country, position, and shoot, country seems to be the most informative. |

The data appears to be in a satisfactory state, and we can proceed with the construction of a regression model.

## 5.2 Regression Analysis

We start by making a model without any data. Then, we choose one that has all the IVs we are studying now[1]. Also, we will make two models: one that doesn’t have categories[2] and another that looks at how IVs interact with each other[3]. In the latter scenario, we shall consider all pairwise interactions and second-order terms.

nullModel = lm(avg\_points ~ 1, data=df)  
# the largest model we will accept  
fullModel1 = lm(avg\_points ~ ., data = df |> select(-Name)) # All variables  
fullModel2 = lm(avg\_points ~ ., data = df |> select(-c(Name, position, country, shoot))) # All numeric variables  
fullModel3 = lm(avg\_points ~ . ^ 2, data = df |> select(-Name)) # All variables with all combinations of interaction

We will utilize a vector function to iterate through the models. A stepwise regression strategy that uses both forward selection and backward elimination will be used.

model = lapply(list(fullModel1, fullModel2, fullModel3), function(x)  
 step(  
 x,  
 scope = list(lower = nullModel, upper = x),  
 direction = "forward"  
 ))

model\_residual = lapply(model,fortify)  
lapply(model,function(x)x$call)

[[1]]  
lm(formula = avg\_points ~ position + country + height + weight +   
 penalty\_dur + g\_diff + sog + age + shoot + duration, data = select(df,   
 -Name))  
  
[[2]]  
lm(formula = avg\_points ~ height + weight + penalty\_dur + g\_diff +   
 sog + age + duration, data = select(df, -c(Name, position,   
 country, shoot)))  
  
[[3]]  
lm(formula = avg\_points ~ (position + country + height + weight +   
 penalty\_dur + g\_diff + sog + age + shoot + duration)^2, data = select(df,   
 -Name))

The interaction model has too many variables, according to the model equations. It is difficult to visualize such a model. Therefore, we will only graph the coefficients of the other two models..

multiplot(model[1:2])

|  |
| --- |
| Figure 6: The coefficients for the full model[1] and the model without categories[2] are shown in the figure. The intercepts and a few categories of the country variable are evidently devoid of statistical significance, as their confidence interval contains zero. The t-test statistic can be used to determine the significance of variables with coefficients close to 0. |

model\_summary = lapply(1:3, function(x)  
 glance(model[[x]]) |>  
 pivot\_longer(everything(),names\_to = "Parameter", values\_to = as.character(x)))  
model\_summary[[1]] |>  
 left\_join(model\_summary[[2]]) |>  
 left\_join(model\_summary[[3]]) |>   
 kable()

Joining with `by = join\_by(Parameter)`  
Joining with `by = join\_by(Parameter)`

Table 4: The table compares the final model parameters for the full model[1], the model without categories[2], and the complex model[3] with all pairs of interactions. Model 3 best explains our dataset, with a R2 value of 0.93. It also has the lowest RSS number under deviance. Then comes the full model[1], which has a R2 of 0.74. All the three models have significant p-values from the F-Test. Complex model 3 is the optimal model for AIC, whereas model 2 is the optimal model for BIC.

| Parameter | 1 | 2 | 3 |
| --- | --- | --- | --- |
| r.squared | 0.7432951 | 0.6950412 | 0.9355846 |
| adj.r.squared | 0.7181827 | 0.6840940 | 0.8057924 |
| sigma | 0.2507203 | 0.2654512 | 0.2081320 |
| statistic | 29.5986881 | 63.4900886 | 7.2083278 |
| p.value | 0.0000000 | 0.0000000 | 0.0000000 |
| df | 18.0000000 | 7.0000000 | 135.0000000 |
| logLik | 2.7635991 | -14.7197348 | 143.0949289 |
| AIC | 34.4728019 | 47.4394696 | -12.1898578 |
| BIC | 100.7369214 | 77.2583234 | 441.7193613 |
| deviance | 11.5663652 | 13.7405447 | 2.9023671 |
| df.residual | 184.0000000 | 195.0000000 | 67.0000000 |
| nobs | 203.0000000 | 203.0000000 | 203.0000000 |

Given the model parameters, we believe it is ideal to proceed with model 1 to avoid both under-fitting and over-fitting.

tidy(model[[1]]) |> mutate(signific = case\_when(  
 between(p.value, 0.05, 0.1) ~ '.',  
 between(p.value, 0.01, 0.05) ~  
 '\*',  
 between(p.value, 0.001, 0.01) ~  
 '\*\*',  
 between(p.value, 0, 0.001) ~  
 '\*\*\*',  
 TRUE ~ ""  
)) |> mutate\_if(is.numeric, ~  
 round(.x, digits = 4)) |> kable()

Table 5: Model 1 includes the dependent variable average points and several predictor variables, such as position, country, height, weight, penalty\_dur, g\_diff, sog, age, shoot, and duration. The position variable is marginally significant. However, the variables country, g\_diff, sog, and duration are highly significant, with p-values below 0.05.

| term | estimate | std.error | statistic | p.value | signific |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | -0.1230 | 0.5860 | -0.2100 | 0.8339 |  |
| positionF | 0.0835 | 0.0439 | 1.9025 | 0.0587 | . |
| countryCzech Republic | 0.2825 | 0.0913 | 3.0952 | 0.0023 | \*\* |
| countryFinland | 0.1381 | 0.0840 | 1.6444 | 0.1018 |  |
| countryFrance | 0.5005 | 0.0972 | 5.1474 | 0.0000 | \*\*\* |
| countryGermany | 0.2785 | 0.0876 | 3.1799 | 0.0017 | \*\* |
| countryHungary | 0.3760 | 0.0877 | 4.2854 | 0.0000 | \*\*\* |
| countryJapan | 0.2985 | 0.0988 | 3.0224 | 0.0029 | \*\* |
| countrySweden | 0.2330 | 0.0857 | 2.7178 | 0.0072 | \*\* |
| countrySwitzerland | 0.3885 | 0.0976 | 3.9792 | 0.0001 | \*\*\* |
| countryUS | 0.1213 | 0.0822 | 1.4752 | 0.1419 |  |
| height | -0.0030 | 0.0040 | -0.7376 | 0.4617 |  |
| weight | 0.0026 | 0.0038 | 0.6907 | 0.4906 |  |
| penalty\_dur | 0.0040 | 0.0043 | 0.9234 | 0.3570 |  |
| g\_diff | 0.0644 | 0.0068 | 9.4051 | 0.0000 | \*\*\* |
| sog | 0.0242 | 0.0039 | 6.1391 | 0.0000 | \*\*\* |
| age | 0.0026 | 0.0046 | 0.5642 | 0.5733 |  |
| shootR | 0.0588 | 0.0406 | 1.4479 | 0.1493 |  |
| duration | 0.0000 | 0.0000 | 3.4323 | 0.0007 | \*\*\* |

Upon completion of the model fitting process, we evaluate the residuals of model 1 for model validation.

m.res = lapply(model\_residual, function(x) x |> select(.fitted, .stdresid))  
model\_residual\_final = m.res[[1]] |> mutate(model=1) |> add\_row(m.res[[2]] |> mutate(model=2)) |> add\_row(m.res[[3]] |> mutate(model=3))  
model\_residual\_final |>  
 ggplot(aes(x = .fitted, y = .stdresid)) +  
 geom\_point() +  
 geom\_hline(yintercept = 0) +  
 stat\_smooth(geom = "smooth") +  
 labs(x = "Fitted Values", y = "Residuals", color="Country") +  
 facet\_wrap(model~.,ncol = 1 )

|  |
| --- |
| Figure 7: The scatter plots of the residual vs. fitted value for the three models. There is heteroscedasticity in all three figures |

model\_residual[[1]] |>  
 ggplot(aes(x = .fitted, y = .resid)) +  
 geom\_point() +  
 geom\_hline(yintercept = 0) +  
 geom\_smooth(se = FALSE) +  
 labs(x = "Fitted Values", y = "Residuals") +  
 facet\_wrap(. ~country, scales = "free\_x", ncol = 3)

|  |
| --- |
| Figure 8: The residual plot with countries shows an approximate homoscedastic behavior when separated out. |

The residual analysis ensures that the model assumptions are met. Additionally, we check the normality of residuals using a QQ plot.

model\_residual\_final |>   
 ggplot(aes(sample=.stdresid)) +  
 stat\_qq() +  
 geom\_abline() +  
 facet\_wrap(model~.,ncol = 2 )

|  |
| --- |
| Figure 9: The figure depicts the qq-plot for the standardized residuals of all three models. All three models show a normal distribution of residuals. |

The QQ plot confirms the normal distribution of residuals. The regression analysis provides insights into the relationships between variables and their impact on average points in women’s ice hockey.

### 5.2.1 Interpretation

From the [Table 4](#tbl-summary-parameters) and [Table 5](#tbl-summary-coefficients), we can interpret the Model 1[[2]](#footnote-3). The very low p-value (< 2.2e-16) for the f-test of the model suggests that the overall model is statistically significant. This implies that there exists a significant linear relationship between the predictors and the response variable.

#### 5.2.1.1 Coefficients:

* **Intercept**: The estimated avg\_points when all other predictors are zero is approximately -0.123. As such this is insignificant.
* **positionF**: Forwards have an estimated increase in avg\_points over the baseline set by the defenders, by approximately 0.0835 when other variables are held constant. This result is marginally significant (*p = 0.0587*).
* **countryCzech Republic, countryFinland, countryFrance, countryGermany, countryHungary, countryJapan, countrySweden, countrySwitzerland, countryUS**: These coefficients represent the estimated change in avg\_points for players from these countries compared to the baseline set by **Canada**.
* **height**: For a one cm increase in height, there is an estimated decrease of approximately 0.0030 in avg\_points. However, this change is not statistically significant, as the p-value is 0.4617.
* **weight**: For a one Kg increase in weight, there is an estimated increase of approximately 0.0026 in avg\_points. However, this change is not statistically significant, as the p-value is 0.4906.
* **penalty\_dur**: For a one second increase in penalty duration, there is an estimated increase of approximately 0.0040 in avg\_points. However, this change is not statistically significant, as the p-value is 0.3570.
* **g\_diff**: For a one-unit increase in goal difference, there is an estimated increase of approximately 0.0644 in avg\_points. This change is highly statistically significant (p < 0.001), suggesting a strong positive relationship with avg\_points.
* **sog**: For a one-unit increase in shots on goal, there is an estimated increase of approximately 0.0242 in avg\_points. This change is highly statistically significant (p < 0.001), indicating a positive association with avg\_points.
* **age**: For a one-year increase in age, there is an estimated increase of approximately 0.0026 in avg\_points. However, this change is not statistically significant, as the p-value is 0.5733.
* **shootR**: Right shooters have an estimated increase in avg\_points over the baseline set by the left shooters, by approximately 0.05878 when other variables are held constant. However, this change is not statistically significant, as the p-value is 0.1493.
* **duration**: For a one second increase in duration, there is an estimated increase of approximately 0.00004 in avg\_points. This change is statistically significant (p = 0.0007), suggesting a positive relationship with avg\_points.

#### 5.2.1.2 Significance:

* The significance of each coefficient is indicated by the p-values. For instance:
  + countryCzech Republic, countryFrance, countryGermany, countryHungary, countryJapan, countrySweden, countrySwitzerland, g\_diff, sog, duration are highly significant (*p < 0.05*).
  + positionF is marginally significant (*p = 0.0587*).
  + height, weight, penalty\_dur, age, shootR are not statistically significant based on conventional significance levels (*p > 0.05*).

#### 5.2.1.3 Model Fit:

* **Multiple R-squared**: The model explains approximately 74.33% of the variance in the dependent variable avg\_points.
* **Adjusted R-squared**: This is a modified version of the R-squared that adjusts for the number of predictors in the model. It is 71.82% in this model.
* **Residual standard error**: This is an estimate of the standard deviation of the residuals. It is 0.2507 in this model.

In summary, the model appears to be a good fit, explaining a substantial portion of the variance in avg\_points. Several predictors are statistically significant, while others are not. Consideration can be given to the practical significance of the coefficients and potential refinement of the model based on domain knowledge. This could potentially hold significant value in incorporating interaction without overfitting.

We also wished to check the semi-partial correlation to understand the unique contribution of each individual predictor variable to the dependent variable (avg\_points), while accounting for the effects of all other predictors. It helps answer the question, “How much does the correlation between a specific predictor and the dependent variable change when we control for the other predictors in the model?” However, we faced issues calculating the same with categorical variables. Hence, we only checked the semi-partial correlation regarding the numeric variables [Table 6](#tbl-spcor).

ppcor::spcor(df |> select\_if(is.numeric), method = "pearson")$estimate |>  
 as\_tibble(rownames = "rowname") |>  
 filter(rowname == "avg\_points") |> select("g\_diff", "sog", "duration") |> kable()

Table 6: The semi-partial correlations of the significant numerical variables.

| g\_diff | sog | duration |
| --- | --- | --- |
| 0.3406532 | 0.3033284 | 0.1112964 |

## 5.3 Applications of the model

We elucidated the linear relationships between the predictor variables and response variables with our model. However, we also wished to explore the business aspects of this model. Therefore, in this section, we try to model an increase in attacking strategy. We implement this in the model through the shots on goal variable. We will assume that all the countries have one game with a hypothetical opponent, and we will calculate the overall points scored by the country. This will be our prediction. We will compare this with the overall observed and estimated points.

pract\_app = df |> group\_by(country) |> arrange(country)  
pract\_app$estimate = model[[1]] |> predict(pract\_app |> select(-c(Name, avg\_points)))  
pract\_app = pract\_app |> mutate(sog = sog \* 1.2)  
pract\_app$prediction = model[[1]] |> predict(pract\_app |> select(-c(Name, avg\_points)))  
pract\_app = pract\_app |> group\_by(country) |> summarise(  
 Observation = sum(avg\_points),  
 Estimate = sum(estimate),  
 Prediction = sum(prediction),  
 "% Increase" = round((Prediction - Observation) / Observation \* 100,2)  
)   
pract\_app |> kable()

Table 7: The table shows how much a country’s overall points go up when their players increase their attacking strategy, which is reflected by a 20% increase in the shots on goal predictor variable. We assume a single game per country in which all players participated. The aggregate points from both observations and estimations are the same, as the country plays a role in the models’ prediction process. We can see that more points are being scored, with an average increase of 12.4%.

| country | Observation | Estimate | Prediction | % Increase |
| --- | --- | --- | --- | --- |
| Canada | 11.144 | 11.144 | 12.724018 | 14.18 |
| Czech Republic | 6.216 | 6.216 | 7.047080 | 13.37 |
| Finland | 15.430 | 15.430 | 16.855398 | 9.24 |
| France | 3.500 | 3.500 | 3.973522 | 13.53 |
| Germany | 6.000 | 6.000 | 6.657133 | 10.95 |
| Hungary | 4.667 | 4.667 | 5.213000 | 11.70 |
| Japan | 2.669 | 2.669 | 3.219832 | 20.64 |
| Sweden | 8.429 | 8.429 | 9.303567 | 10.38 |
| Switzerland | 5.572 | 5.572 | 6.175982 | 10.84 |
| US | 15.773 | 15.773 | 17.203230 | 9.07 |

pract\_app |> pivot\_longer(cols = Estimate:Prediction) |>  
 ggplot(aes(x = reorder(country, Observation), y = value)) +  
 geom\_col(aes(fill = name), position = "dodge") +  
 geom\_text(  
 aes(label = ifelse(  
 name == "Estimate", paste(`% Increase`, "%"),  
 ""  
 )),  
 position = position\_dodge(width = .9),  
 vjust = 0,  
 show.legend = FALSE  
 ) +  
 theme(legend.title = element\_blank()) +  
 labs(x="Country", y="Aggregate Points")

|  |
| --- |
| Figure 10: The figure compares the relative increase in prediction regarding the estimate of the points scored by each country when their players increase their attacking strategy, which is reflected by a 20% increase in the shots on goal predictor variable. As per the model, there is no expense associated with enhancing the offensive approach. |

We can see that more points are being scored, with an average increase of 12.4% across the countries. As per the model, there is no expense associated with enhancing the offensive approach.

# References

“Basic Ice Hockey Rules Sportsa Articles.” n.d. Accessed October 2, 2023. <https://sportsa.com/articles/basic-ice-hockey-rules.html>.

“Canada Player Statistics.” 2023. April 16, 2023. <https://www.iihf.com/pdf/498/ihw4980can_83_8_0>.

“Czechia Player Statistics.” 2023. April 16, 2023. <https://www.iihf.com/pdf/498/ihw4980cze_83_8_0>.

“Finland Player Statistics.” 2023. April 16, 2023. <https://www.iihf.com/pdf/498/ihw4980fin_83_7_0>.

“France Player Statistics.” 2023. April 16, 2023. <https://www.iihf.com/pdf/498/ihw4980fra_83_4_0>.

“Germany Player Statistics.” 2023. April 16, 2023. <https://www.iihf.com/pdf/498/ihw4980ger_83_6_0>.

“Hungary Player Statistics.” 2023. April 16, 2023. <https://www.iihf.com/pdf/498/ihw4980hun_83_6_0>.

“IIHF World Championship (Women) 2023 - Player Stats.” n.d. QuantHockey. Accessed August 30, 2023. <https://www.quanthockey.com/whc-women/en/seasons/whc-women-players-stats.html>.

“Japan Player Statistics.” 2023. April 16, 2023. <https://www.iihf.com/pdf/498/ihw4980jpn_83_7_0>.

Nigg, Claudio R., Andreas Gessner, Carina Nigg, Marco Giurgiu, and Rainer Neumann. 2020. “Demographic, Physiological, Psychological, and on-Ice Performance Indicators Predict Plus/Minus Status of Recreational Ice Hockey Players Across a Season.” *German Journal of Exercise and Sport Research* 50 (4): 463–69. <https://doi.org/10.1007/s12662-020-00679-2>.

“Sweden Player Statistics.” 2023. April 16, 2023. <https://www.iihf.com/pdf/498/ihw4980swe_83_9_0>.

“Switzerland Player Statistics.” 2023. April 16, 2023. <https://www.iihf.com/pdf/498/ihw4980sui_83_9_0>.

“USA Player Statistics.” 2023. IIHF International Ice Hockey Federation. April 16, 2023. <https://www.iihf.com/pdf/498/ihw4980usa_83_8_0>.

“Women’s World Championship Statistics.” n.d. IIHF International Ice Hockey Federation. Accessed August 30, 2023. <https://www.iihf.com/en/events/2023/ww/teamstats/scoringefficiency>.

1. The dataset is provided in <data.csv> [↑](#footnote-ref-2)
2. Refer to [Figure 6](#fig-coefficient-plot) for coefficient plot. [↑](#footnote-ref-3)