Assignment 2: Quantifying NHL Team Strength

Due September 26, 2025

In this assignment you will quantify team strength using National Hockey League (NHL) data from last season. First, you will calculate Pythagorean expectation and see if it is a better predictor of future success than points percentage. Then, you will fit a Bradley-Terry model and compare differences in how team strength is quantified.

Submit your assignment to Gradescope as a PDF file generated from your R Markdown. Make sure your PDF includes all code, outputs, and written answers to questions. All code and outputs used to answer the written questions must be included in the PDF for full credit.

Loading the data

You will use data from the R package `nhlscraper` in this assignment. This package scrapes data from the NHL API.

```
# Run line of code below once to install nhlscraper
# install.packages('nhlscraper')
library(nhlscraper) # scrapes data from NHL API
library(tidyverse)
```

Run the code below to load in game-level data from the NHL 2024-25 season.

```
# only consider game dates from 2024-25 season
season_days <- seq.Date(
  from = as.Date("2024-10-04"),
  to   = as.Date("2025-04-17"),
  by   = "day"
)
# convert to string format
season_days_str <- as.character(season_days)
# get game-level stats for regular season games
team_game_stats <- get_team_statistics(season = get_season_now()$seasonId,
                                       report = "summary",
                                       is_aggregate = FALSE,
                                       is_game = TRUE,
                                       dates = season_days_str,
                                       game_types = 2)
```

Pythagorean Expectation [15 points]

1 (5 points). In the NHL, a win is worth 2 points, an overtime loss (OTL) is worth 1 point and a regulation loss is worth 0 points. A team's points percentage (P%) is calculated as:

$$P\% = \frac{\text{Total Points}}{\text{Total Possible Points}}$$

where "Total Possible Points" is equal to the number of games played times two. For a full regular season, this equals 82 x 2 = 164.

Summarize `team_game_stats` to calculate each team's points percentage and Pythagorean expectation (using goals for and goals against, with $\alpha = 2.11$) for the season.

Which team most over-performed and which team most under-performed their expectation, based on the difference between Pythagorean expectation and points percentage?

```
alpha <- 2.11

team_summary <- team_game_stats %>%
  group_by(teamFullName) %>%
  summarise(
    goals_for = sum(goalsFor),
    goals_against = sum(goalsAgainst),
    wins = sum(wins),
    otl = sum(otLosses),
    games_played = n()
  ) %>%
  mutate(
    total_points = 2 * wins + otl,
    total_possible = games_played * 2,
    points_pct = total_points / total_possible,
    pyth_exp = (goals_for ^ alpha) / ((goals_for ^ alpha) + (goals_against ^ alpha)),
    diff = points_pct - pyth_exp) %>%
  arrange(desc(diff))

# Which team over-performed and under-performed:
team_summary %>%
  filter(diff == max(diff) | diff == min(diff))
```

```
## # A tibble: 2 x 11
##   teamFullName     goals_for goals_against  wins   otl games_played total_points
##   <chr>                <int>         <int> <int> <int>        <int>        <dbl>
## 1 Calgary Flames         220           236    41    14           82           96
## 2 Tampa Bay Light~       292           216    47     8           82          102
## # i 4 more variables: total_possible <dbl>, points_pct <dbl>, pyth_exp <dbl>,
## #   diff <dbl>
```

The team that over-performed was the Calgary Flames. The team that under-performed was the Tampa Bay Lightning.

2 (5 points). Create two new dataframes called `first_half` and `second_half` that filter `team_game_stats` to game dates before 1/05/2025, and 1/05/2025 and later, respectively. For both datasets calculate Pythagorean expectation (using $\alpha = 2.11$) and points percentage.

Note: The number of games in each half will be different from a full season (82 games). Make sure you use the actual number of games played in each half when calculating percentages.

```
first_half <- team_game_stats %>%
  filter(gameDate < "2025-01-05") %>%
  group_by(teamFullName) %>%
  summarise(
    goals_for = sum(goalsFor),
    goals_against = sum(goalsAgainst),
    wins = sum(wins),
    otl = sum(otLosses),
    games_played = n()
  ) %>%
```

```
  mutate(
    total_points = 2 * wins + otl,
    total_possible = games_played * 2,
    points_pct = total_points / total_possible,
    pyth_exp = (goals_for ^ alpha) / ((goals_for ^ alpha) + (goals_against ^ alpha))
  )

second_half <- team_game_stats %>%
  filter(gameDate >= "2025-01-05") %>%
  group_by(teamFullName) %>%
  summarise(
    goals_for = sum(goalsFor),
    goals_against = sum(goalsAgainst),
    wins = sum(wins),
    otl = sum(otLosses),
    games_played = n()
  ) %>%
  mutate(
    total_points = 2 * wins + otl,
    total_possible = games_played * 2,
    points_pct = total_points / total_possible,
    pyth_exp = (goals_for ^ alpha) / ((goals_for ^ alpha) + (goals_against ^ alpha))
  )
```

3 (5 points). Join `first_half` and `second_half` on team. Calculate the correlation between first half Pythagorean expectation and second half points percentage. Calculate the correlation between first half points percentage and second half points percentage. Describe and interpret your results.

```
combined <- first_half %>%
  select(teamFullName, points_pct_first = points_pct, pyth_first = pyth_exp) %>%
  inner_join(
    second_half %>% select(teamFullName, points_pct_second = points_pct, pyth_second = pyth_exp), by =

cor_pyth_to_second <- cor(combined$pyth_first, combined$points_pct_second)
cor_points_to_second <- cor(combined$points_pct_first, combined$points_pct_second)

cor_results <- data.frame(
  Comparison = c("First-half Pythagorean vs Second-half Points %",
                 "First-half Points % vs Second-half Points %"),
  Correlation = c(round(cor_pyth_to_second, 3),
                  round(cor_points_to_second, 3))
)

cor_results
```

```
##                                       Comparison Correlation
## 1 First-half Pythagorean vs Second-half Points %       0.669
## 2    First-half Points % vs Second-half Points %       0.643
```

The correlation between first-half Pythagorean expectation and second-half points percentage is 0.669, suggesting that underlying goal-based performance is a fairly good predictor of future results. The correlation between first-half points percentage and second-half points percentage is 0.643. Since it is slightly weaker, we can conclude that raw win-loss performance is slightly less predictive than goal-based metrics. Overall, this supports the idea that Pythagorean expectation captures true team strength better than actual points percentage, which may be influenced by luck or close-game outcomes.

Bradley-Terry model [15 points]

Next, you will fit a Bradley-Terry model to quanitfy team strength. Run the following code to get a reformatted game result dataset, called `bt_df`, that can be used to fit the model.

```
# Reformat dataset
bt_df <- team_game_stats %>%
  mutate(value = ifelse(homeRoad == "H", 1, -1)) %>%
  mutate(home_team_win = wins[homeRoad == "H"][1], .by = gameId) %>%
  select(gameId, teamFullName, value, home_team_win) %>%
  pivot_wider(names_from = teamFullName, values_from = value, values_fill = 0) %>%
  arrange(gameId)
```

4 (5 points). Using `bt_df`, fit a Bradley-Terry model, with `home_team_win` as the response and the team indicator columns as predictors. Output a summary of the model. According to the model, what is the strongest and weakest NHL team that season? Does this agree with the strongest and weakest team according to Pythagorean expectation and points percentage?

```
# Fit Bradley-Terry model
bt_model <- glm(home_team_win ~ . ,
                data = bt_df %>% select(-gameId),
                family = binomial())

sort(summary(bt_model)$coefficients[-1,1],decreasing = TRUE)
```

```
##         `Winnipeg Jets`   `Toronto Maple Leafs`   `Washington Capitals`
##              0.80191553              0.58136979              0.52242211
##  `Vegas Golden Knights`          `Dallas Stars`    `Colorado Avalanche`
##              0.47166286              0.46892544              0.41650548
##     `Los Angeles Kings`       `Edmonton Oilers`   `Tampa Bay Lightning`
##              0.35823806              0.35803478              0.32998310
##      `Florida Panthers`   `Carolina Hurricanes`        `Minnesota Wild`
##              0.32469284              0.32340550              0.21669037
##       `Ottawa Senators`       `St. Louis Blues`     `New Jersey Devils`
##              0.21661110              0.17648116              0.07240208
##    `Montréal Canadiens` `Columbus Blue Jackets`     `Detroit Red Wings`
##             -0.01814313             -0.02715635             -0.06719652
##      `New York Rangers`      `Utah Hockey Club`     `Vancouver Canucks`
##             -0.08017425             -0.12449449             -0.14907234
##        `Buffalo Sabres`         `Anaheim Ducks`    `New York Islanders`
##             -0.22606913             -0.27364031             -0.28402326
##        `Seattle Kraken`   `Pittsburgh Penguins`         `Boston Bruins`
##             -0.29252553             -0.33008566             -0.36510462
##   `Philadelphia Flyers`   `Nashville Predators`    `Chicago Blackhawks`
##             -0.38087449             -0.53440637             -0.81574243
##       `San Jose Sharks`
##             -1.12861968
```

The strongest NHL team is the Winnipeg Jets. The weakest team is San Jose Sharks. This does not agree with the strongest and weakest team according to Pythagorean expectation and points percentage which are the Calgary Flames(strongest) and Tampa Bay Lightning(weakest) respectively.

5 (4 points). Interpret the intercept of the Bradley-Terry model. What is the effect of home ice advantage?

```
exp(coef(bt_model)[1])
```

```
## (Intercept)
##    1.312446
```

The home team is 1.3 times more likely to win because of home advantage by playing on their own ice.

6 (3 points). In the 2025 Stanley Cup final, the Florida Panthers played the Edmonton Oilers. According to the Bradley-Terry model what's the probability that the Panthers beat the Oilers in a game?

```
new_game <- tibble(
  `Florida Panthers` = 1,
  `Edmonton Oilers` = -1
)
all_teams <- setdiff(names(bt_df), c("gameId", "home_team_win"))
for(team in all_teams) {
  if(!team %in% names(new_game)) new_game[[team]] <- 0
}

# Use model to predict
prob_panthers_win <- predict(bt_model, newdata = new_game, type = "response")
prob_panthers_win*100
```

```
##        1
## 55.93565
```

The probability that the Panthers beat the Oilers in a game is approximately 55.9%.

7 (3 points). Using the probability of the Panthers winning that you calculated in the previous question, calculate the probability that the Panthers win the Stanley cup final, which is a best-of-7 series. To win a best-of-7 series, the Panthers must win 4 games before the Oilers do.

Hint: to calculate the probability of winning a best-of-7 series, you can add together the probability of winning the series in 4 games, 5 games, 6 games and 7 games:

- To win in 4 games the team must win the first 4 games.
- To win in 5 games, the team must win 3 of the first 4 games and the 5th game.
- To win in 6 games, the team must win 3 of the first 5 games and the 6th game.
- To win in 7 games, the team must win 3 of the first 6 games and the 7th game.

```
p <- prob_panthers_win

series_win_prob <- p^4 + 4*p^4*(1-p) + 10*p^4*(1-p)^2 + 20*p^4*(1-p)^3
series_win_prob*100
```

```
##       1
## 62.8028
```

The probability of the Panthers winning a best-of-7 series against the Oilers for the Stanley Cup is 62.8%