Using NHL Regular Season Result Data to Predict Playoff Outcomes

Trevor Garrood and Evan Bodenstab

1. We are going to answer the following research questions:
    a. Can Pythagorean expectation accurately predict the amount of wins a team gets in the regular season?

        Synopsis: We want to determine if Pythagorean expectation (number of points for divided by number of points for/against) can accurately predict the number of regular season wins/playoff performance. This method has proven to work well for college and professional basketball, so we will determine if it can be applied to the NHL as well.

        Results: Based on our results, pythagorean expectation can be used very successfully in determining an estimate of how many wins a team got in the regular season of any given year.

    b. Do regular season win streaks correlate with overall season performance?

        Synopsis: Among numerous NHL enthusiasts, the go to method of predicting total performance (Win-Loss-OT record) is to use the frequency and average length of regular season win streaks. We will calculate these stats for every team throughout a season, and then determine if there is reasonable statistical evidence to conclude that win-streaks do connect with better performances in the regular season.

        Results: We analyzed how both the number of win streaks and the longest total win streak for each team in various seasons correlate to how they performed in that season, and we determined that win streaks are an extremely poor indicator of whether or not a team's regular season is very successful.

    c. How does playing a longer season affect overall season performance the next year?

        Synopsis: When a team does well in the playoffs, their season can last multiple months longer than teams who do not make the playoffs at all. We will measure if playing for a longer amount of time affects a team's performance the next year. Some experts say that it can help a team stay in shape for the next year, but others say that it wears the team down, and we want to find a statistical conclusion.

        Results: We found there to be some moderate correlation between the length of a team's previous year and the number of wins they would be able to get the following year in the regular season.

d. Can a team's regular season statistics be used to predict playoff outcomes?

Synopsis: The combined stats of total number of wins, frequency/length of win streaks, and total number of goals scored throughout the regular season provide a pretty large indicator of a team's potential playoff performance. We want to train a machine learning model to try and predict playoff outcomes and see if these factors (or a variation of them) can accurately predict future playoff outcomes.

Results: Our machine learning model was able to predict playoff performances with around 80% accuracy, making it mostly successful in using regular season data to predict playoff success across all of the teams.

2. Motivation and Background:

The results of NHL playoff series can often be unexpected. Most people might use regular season win-loss records as the only indicator of how well they expect a team to perform, which can lead to surprises as playoff series outcomes often do not follow the simple pattern of regular season win percentage. We want to take a deeper look at regular season statistics for every team to see if other statistics besides simple win-loss ratios might better predict how well any team will do in the playoffs. Many NHL enthusiasts predict postseason performance based on frequency and average length of win streaks for any team throughout the season. We are going to look at patterns of regular season win streaks to see if these can better indicate how well a team will do in the playoffs. We also want to see if the length of a team's postseason run has any weight in that team's performance the following season. Some experts say an extended season that goes into playoffs can help a team stay in shape to be ready for the next season, while others say the longer season can wear teams down. We want to determine whether either of these claims are true, or if some other claim might have more statistical evidence. Finally, we plan on using the aforementioned win streak statistics and trends, as well as goal and win totals, to build up a machine learning model that can predict playoff outcomes to see if some variation of these trends might give accurate insight into future playoff outcomes.

3. Our dataset: https://www.kaggle.com/martinellis/nhl-game-data

We will be using three different datasets from this link. The first is `game_teams_stats.csv` which contains scores of games as well as hits, penalties, and shots. The second is `game.csv` which contains how many goals for home/away as well as faceoff percentage and if the game is playoff or regular season.The third is `team_info.csv` which contains all of the NHL teams along

with some useful identifying information we can use to link teams to the other two datasets.

4. Methodology:

Old:

We will need to calculate a few key results for each team in order to do our analysis. Our data set only reports scores of games, so we will need to determine which team won each game and then add that win to the respective team's total wins for the regular season. We then will need to count the number and duration of win streaks a team has, as well as how many goals that team scores in a season and how many they got scored against them. We also will need to sum up the time duration of every game a team plays to calculate how long their entire season is (playoffs and regular). From there, we will find out how well a team makes it in the playoffs. We will do this by counting how many total wins each team gets, what round they make it to (total wins modulus four because best of seven series), how many goals they score, and how many goals are scored against them. We will use this data to rank each team (1-16) based on the performance data we just calculated (first on total wins, then goals for, and then goals against. Now that all of our statistics have been prepared, we can begin analyzing them. First, we will rank teams based on the total number of win streaks and duration of those streaks and compare it to the results of our playoff rankings, to determine if win streaks are a good indicator of playoff performance. Then, we will rank teams based on the total duration of their season, and determine if that is a good indicator of performance in the next year's playoffs. Then, we will use a teams total wins, goal differential (goals for - goals against), and number of win streaks, and playoff performance, to train a machine learning system to predict team's performance for future years. With the machine learning model, we will then determine if it has any statistical significance and if accurate predictions can be made with the statistics we calculated.

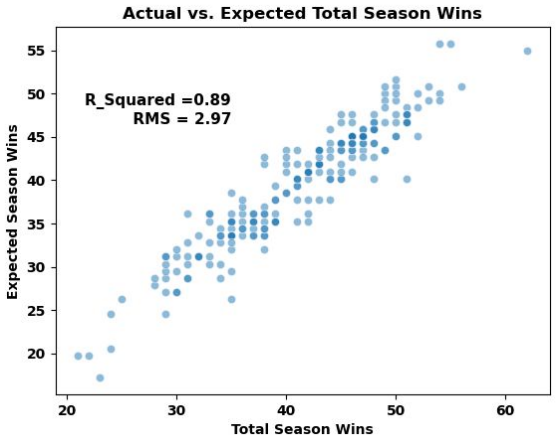
New:

Since we decided to use three datasets, we first merged them together on columns they shared. We then implemented a method that could take our merged dataset or any other data with the necessary columns and create a new dataset containing the stats required to perform our computations. The new dataset is indexed by team for each season, so each team has a new row for each season. Stats for each team across each season include the team's goals for, goals against, total wins in that regular season, number of win streaks, longest win streak, faceoff percentage, penalty minutes, hits, and win-loss ratio. When this dataset was created, we also

had the program perform Pythagorean expectation, which divides the goals for by the goals for plus goals against for each team, to guess how many games the team won for each season. We then used this dataset to analyze other patterns between various statistics and total regular season wins to determine if the categories of stats we chose had any correlation to the number of wins in a season. We implemented a function to plot scatter plots showing correlation between different categories of data, as well as to calculate R-squared for each scatterplot and the root-mean-squared error between the columns given. Below are some of the scatterplots we used to see if trends existed between categories such as number of win streaks in a season and length of the previous season and the total number of wins in the regular season, as well as the scatter plot showing how well our Pythagorean Expectation performed at predicting the total number of wins each team would get. Finally, we used this dataset to train a machine learning model. We used all of the columns excluding data concerning the playoffs as features and had the model predict the number of wins each team would get in the playoffs as labels. The results of runs of this model for two different seasons are also displayed in the next section.

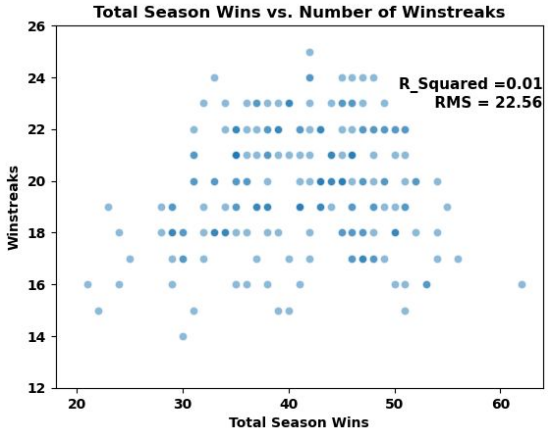
5. Results:
   a. **Can Pythagorean expectation accurately predict the amount of wins a team gets in the regular season?**

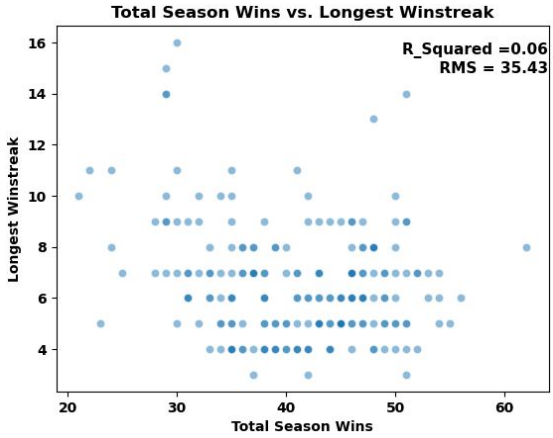
      According to our research, Pythagorean expectation is a viable way of inferring a rough estimate of how many games any team won in any given season. The scatterplot below shows the results of comparing each team's estimated number of wins based on Pythagorean expectation to their actual number of wins in the season. It also has a high r^2 value (correlation) and low RMS error (low error on predictions).



b. Do regular season win streaks correlate with overall season performance?

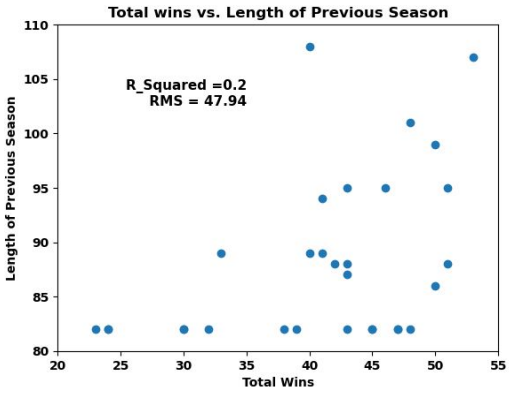
No, they do not. Based on our research, there was almost no correlation between the number of regular season wins and either number of win streaks or longest win streak. This is a bit surprising since many enthusiasts and even experts relate regular season win streaks to strong seasons, but clearly the data shows no connection between the two. Both have low r^2 values (little/no correlation) and very high RMS errors (inaccurate predictors).





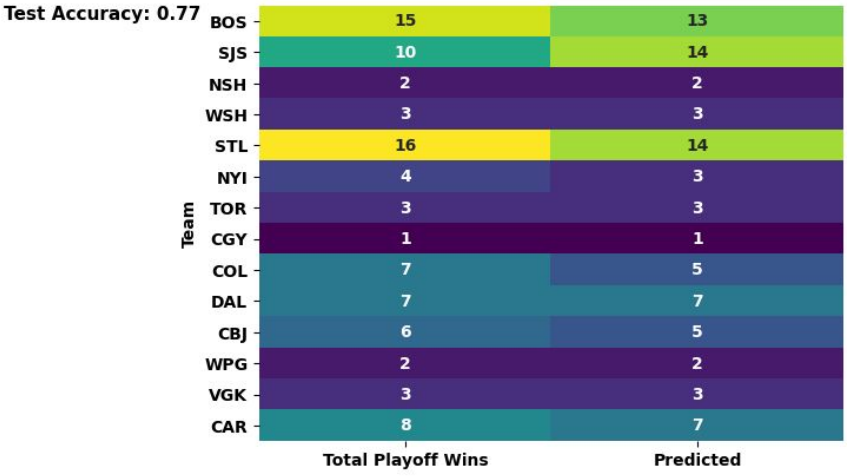
c. How does playing a longer season affect overall season performance the next year?

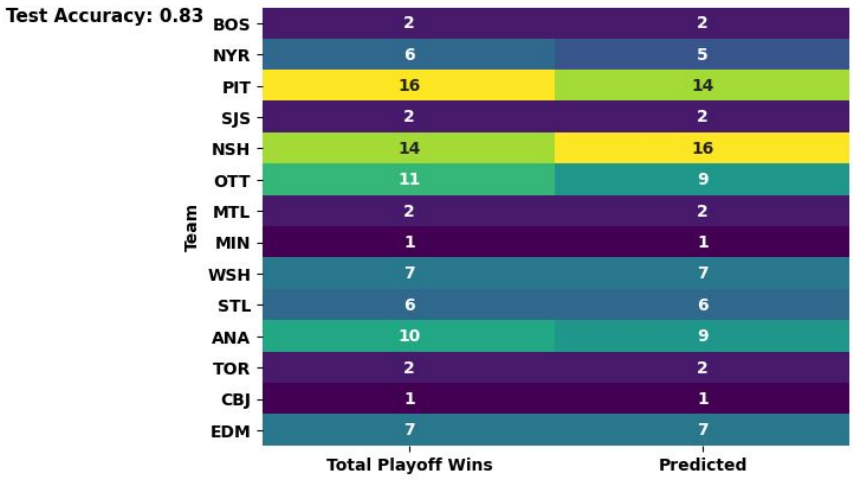
We saw a mild trend in our data that appears to imply that a team having an extended previous season due to a trip to the playoffs might correlate to the team doing better the next season. Based on the scatterplot below, which compares the number of games each team played in the 2013-14 season to the number of wins they got in the 2014-15 season, teams that do not go to the playoffs tend to get less wins than the teams that did play in the playoffs the following year. This result is contrary to what many experts have been saying about a longer season (many experts warn that going longer into the playoffs can produce burnout and decrease wins the next year). Has low r^2 value(little/no correlation) and very high RMS error (inaccurate predictors).



d. Can a team's regular season statistics be used to predict playoff outcomes?

The two charts shown below display the results from our machine learning model when given regular season statistics for each team in the league as features and predicting how each team will do in the playoffs based on those regular season statistics. Our model was able to get 77% and 83% test accuracy with two different sets of test data. Based on these numbers, we believe that regular season stats can be used as an indicator to predict outcomes in playoffs, and they can also be used to make a machine learning model to do so as well.





6. Challenge Goals:

Machine learning:

We created a **machine learning** model that predicts playoff performances for teams based on various regular season stats. The features include statistics such as regular season wins, goals for, goals against, and win streak data among other stats to predict its labels, which are numbers of playoff wins it predicts different teams will have. As shown in the colored charts in our results section, our model performed fairly well, resulting in a test accuracy of 77% and 83% for two different seasons, being the 2016-17 and 2018-19 seasons.

Multiple Perspectives:

We tried to use **multiple perspectives** in analyzing our data and using it to predict other trends in statistics. We tried utilizing a few very different categories of statistics to attempt to predict regular season performance, as displayed in the scatterplots in the results section. We wanted to use some different angles in looking at trends to cover all areas of hockey stats, in case there could be strong correlation in a stat category that no one would ever think of. We also used all of these statistics as features to train our playoff predicting model, using yet another new perspective.

Multiple Datasets:

We think our program effectively combines **multiple datasets**. We decided to use three datasets and merge them together to enable us to use a single, consolidated dataset to answer our questions more easily. We needed both the game stats and team stats per game datasets to gather all

the necessary information produced through each season, and we needed the team information dataset to distinguish stats for each team across the different seasons. By combining our datasets and then creating a new dataset out of the necessary information from the combined one, we were able to structure our data so that it could be used efficiently.

7. Work plan:
    - Calculate all needed data for analysis (wins, win streaks, rankings) **approx. 3hrs**
    - Plot comparisons of different statistics and determine if they are good indicators of positive or negative season performance **approx. 3hrs**
    - Train a machine learning algorithm to predict playoff performance **approx. 4 hrs**
    - We are planning to collaborate using google notebooks to write our code (can share code as well as see outputs for tests) as well as audio/visual communication over facetime or zoom. Our final deliverable is either going to be in the form of a blog post in Microsoft Word because that will allow us to demonstrate our plots as well as give us adequate space to explain our findings
    - Evaluation: Our estimates from our initial work plan were decently accurate, although they don't encompass how much focus we had to put into different tasks. Calculating the necessary data that we used for our analyses definitely took longer than we expected, but that was due to switching datasets after we had originally been planning on using a different one that we decided was inadequate. Once we found the datasets that we decided to use though, processing the data was rather straightforward. Plotting the trends we found did not take as much time as we expected, although we had to rework our code to be more efficient in plotting different categories which took some unplanned time. Designing our machine learning model actually took far less time than we assumed it was, due to our ability to get most of the code we required from past lectures and homeworks and adapt it to do what we needed it to do. Overall, although our work plan did not account for some obstacles that we ended up having to work around, it was fairly accurate in describing what we would do and how long it would take.

8. Testing:

We primarily used Jupyter Notebook for testing our code. This allowed us to visually cross compare stats with the NHL's official stats on their website, which ensures us that the data we calculated is accurate and actually what happened for the respective team and season. We also used calculations on subsections of the larger data file to ensure all calculations gave the correct result. You can trust our results because they are consistent with data directly from the NHL website, as well as data that could not be directly checked was cross-compared with manually computed results from smaller subsections of the dataset

```
""" Tests output of season_data_generator """
season_data_generator(20182019, 'NJD', df)
```

```
{'Expected Playoff Wins': 5.92,
 'Expected Season Wins': 30.34,
 'Faceoff Percentage': 49.41,
 'GA': 278,
 'GF': 219,
 'Hits': 1924,
 'Longest Win Streak': 6,
 'PIM': 745,
 'Playoff Round': 0,
 'Season': 20182019,
 'Team': 'NJD',
 'Total Games': 82,
 'Total Playoff Wins': 0,
 'Total Season Wins': 31,
 'Win Ratio': 0.37,
 'Winstreaks': 21}
```

```
""" Tests output of season_data_generator
    with a team that makes the playoffs  """
season_data_generator(20182019, 'BOS', df)
```

```
{'Expected Playoff Wins': 9.44,
 'Expected Season Wins': 48.38,
 'Faceoff Percentage': 50.65,
 'GA': 217,
 'GF': 257,
 'Hits': 1876,
 'Longest Win Streak': 4,
 'PIM': 797,
 'Playoff Round': 4,
 'Season': 20182019,
 'Team': 'BOS',
 'Total Games': 106,
 'Total Playoff Wins': 15,
 'Total Season Wins': 49,
 'Win Ratio': 0.59,
 'Winstreaks': 22}
```

```
Train Accuracy: 0.972972972972973
Test  Accuracy: 0.84375
Test predictions: [ 7  2  0 11 16  0  0  2  7  5  2
  0 16  2 16  0  4  1 10  0  0  4  1  0  0  0  0  0
 11  7  2  2  0  0  2  0  0  0  0  0  0  0  0  2]
Actual Results: Team
MIN     6
NSH     2
FLA     0
TBL    11
NYR    13
       ..
OTT     0
NJD     0
NSH     0
EDM     0
CBJ     2
Name: Total Playoff Wins, Length: 64, dtype: int64
```

9. Collaboration: We had no outside help on this project outside of course materials and documentation for pandas and matplotlib. We did reference Stack Overflow when attempting to add text to our plots, as that was something we did not cover in class and matplotlib documentation did not cover certain aspects of implementation for our needs.