STA 199 Group 1-3 Final Written Report

due October 15, 2021 by 11:59 PM

Thomas Barker, Phoebe Ducote, Ebenezer Teshome, Alexander Du

October 8 2021

Introduction and Data

Link to the dataset

Introduction

This data set was compiled by fivethirtyeight from hockey-reference.com, a site that has collected statistics for every hockey game since the 1917-1918 NHL season with the help of the NHL's dedicated statisticians. One of the most interesting aspects of this dataset is the elo rating system that fivethirtyeight has implemented. Using the data, fivethirtyeight has calculated each team's elo rating before and after each game. Like in many other games and sports, elo ratings are used to determine the relative skill levels and power rankings of, in this case, teams, and are useful for predicting which team will win a game before it is played. In this report, we investigate which statistics have the greatest impact on the elo rating of the Carolina Hurricanes, a NHL team based in Raleigh, North Carolina.

Each case is a single hockey game, and relevant variables include:

'season' - the NHL season in which a game was played

'home_team' - the team playing at home for a game

'away_team' - the team playing away for a game

'home_team_pregame_rating' - the home team's elo rating before a game was played

'away_team_pregame_rating' - the away team's elo rating before a game was played

'home_team_score' - the number of goals scored by the home team

'away_team_score' - the number of goals scored by the away team

'home_team_postgame_rating' - the home team's elo rating after a game was played

'away_team_postgame_rating' - the away team's elo rating after a game was played

Data Set Manipulation

We filtered the NHL data set for games played by the Carolina Hurricanes, hereafter referred to as the Canes. Because we are interested in the recent and future performance of the Canes, we decided to limit the data set to games played between the 2014-15 season to the 2020-21 season. We believe this is more representative of the current state of the Canes due to the high turnover rate in hockey and other sports. Furthermore, we mutated the data to add several variables to carry out our analysis. The variable 'canes_result' tells us whether the Canes won at home, won away, lost at home, or lost away. The variable 'canes_win' tells us whether the Canes won or lost. The variable 'canes_perc_goals' tells us the percentage of the total goals in

a game that were scored by the Canes. The variable 'canes_net_elo' tells us how much the Hurricanes elo rating changed after each game.

Question and Hypotheses

Question: Which factors and statistics affect the Carolina Hurricanes' elo rating the most?

Hypothesis 1: Winning away increases the Canes' elo more than winning at home.

Hypothesis 2: Losing at home decreases the Canes' elo more than losing away.

Hypothesis 3: When the Canes win, the greater the percentage of total goals scored by the Canes, the greater their elo will increase. When the Canes lose, the greater the percentage of total goals scored by the Canes, the less their elo will decrease.

Methodology

Exploratory Data Analysis

```
## # A tibble: 1 x 4
##    mean median   min   max
##   <dbl>  <dbl> <dbl> <dbl>
## 1  3.16   2.99  1.35  6.75
```

When the Canes won at home, their mean elo increase was 3.16, their median elo increase was 2.99, their minimum elo increase was 1.35, and their maximum elo increase was 6.75.

```
## # A tibble: 1 x 4
##    mean median   min   max
##   <dbl>  <dbl> <dbl> <dbl>
## 1  4.01   3.65  1.41  7.34
```

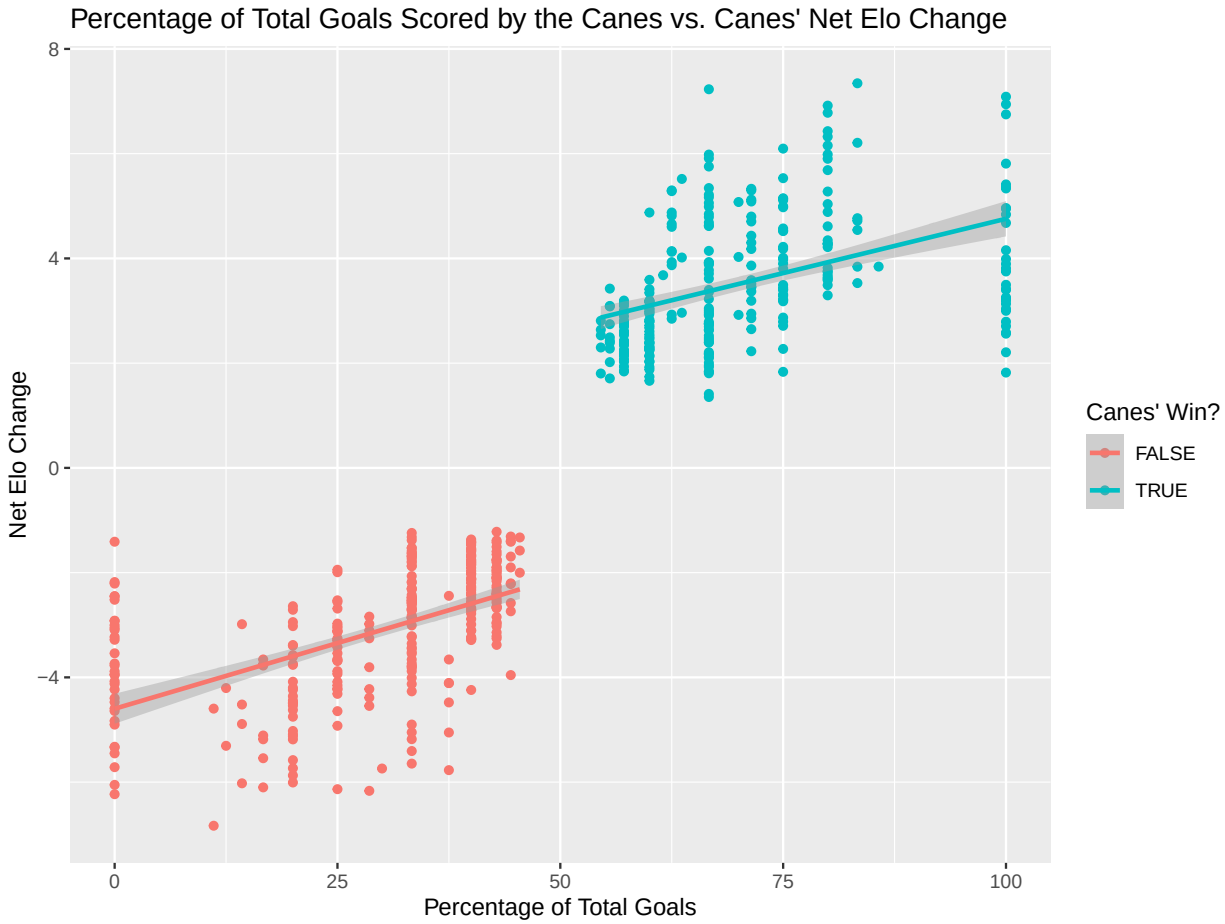
When the Canes won away, their mean elo increase was 4.01, their median elo increase was 3.65, their minimum elo increase was 1.41, and their maximum elo increase was 7.34.

```
## # A tibble: 1 x 4
##    mean median   min   max
##   <dbl>  <dbl> <dbl> <dbl>
## 1 -3.68  -3.35 -6.83 -1.85
```

When the Canes lost at home, their mean elo decrease was 3.68, their median elo decrease was 3.35, their minimum elo decrease was 1.85, and their maximum elo decrease was 6.83.

```
## # A tibble: 1 x 4
##    mean median   min   max
##   <dbl>  <dbl> <dbl> <dbl>
## 1 -2.68  -2.54 -6.23 -1.22
```

When the Canes lost away, their mean elo decrease was 2.68, their median elo decrease was 2.54, their minimum elo decrease was 1.22, and their maximum elo decrease was 6.23.



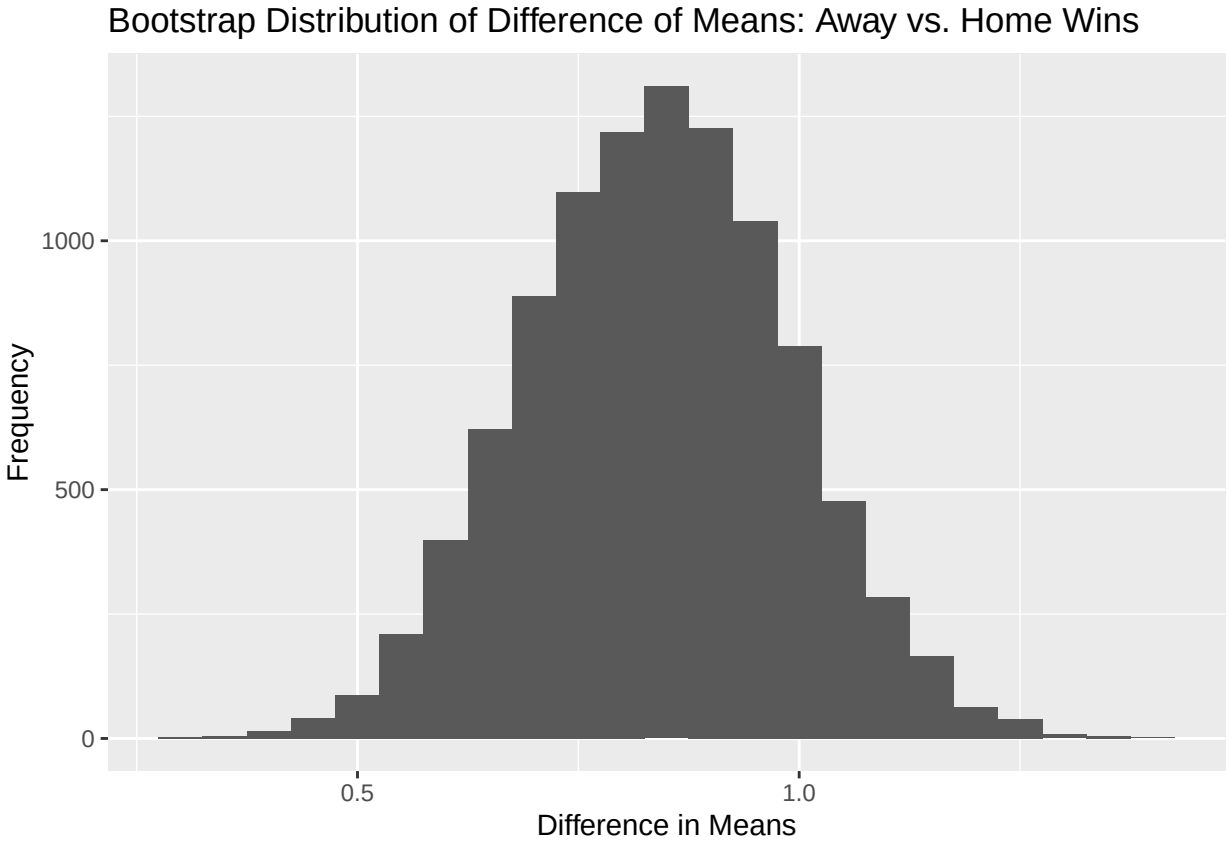
The scatterplot shows a positive relationship between the percentage of total goals scored by the Canes and the Canes' net elo for both wins and losses.

Explanation of Statistical Methods

For our analysis, we use a 95% confidence interval for the difference in mean net elo between the Canes' Home and Away wins. Next, we also use a 95% confidence interval for the difference in mean net elo between the Canes' Home and Away losses. We chose the mean as our sample statistic because it is a good measure of central tendency, especially since the data does not have significant outliers. In addition, we chose a confidence interval because it is an effective way to determine whether or not the mean net elo between the Canes' Home and Away wins is the same or if not, in which direction they differ. Finally, we run two linear regressions, one for the Canes' wins and one for the Canes' losses, to examine the relationship between the percentage of goals scored by the Canes and the Canes' net elo.

Results

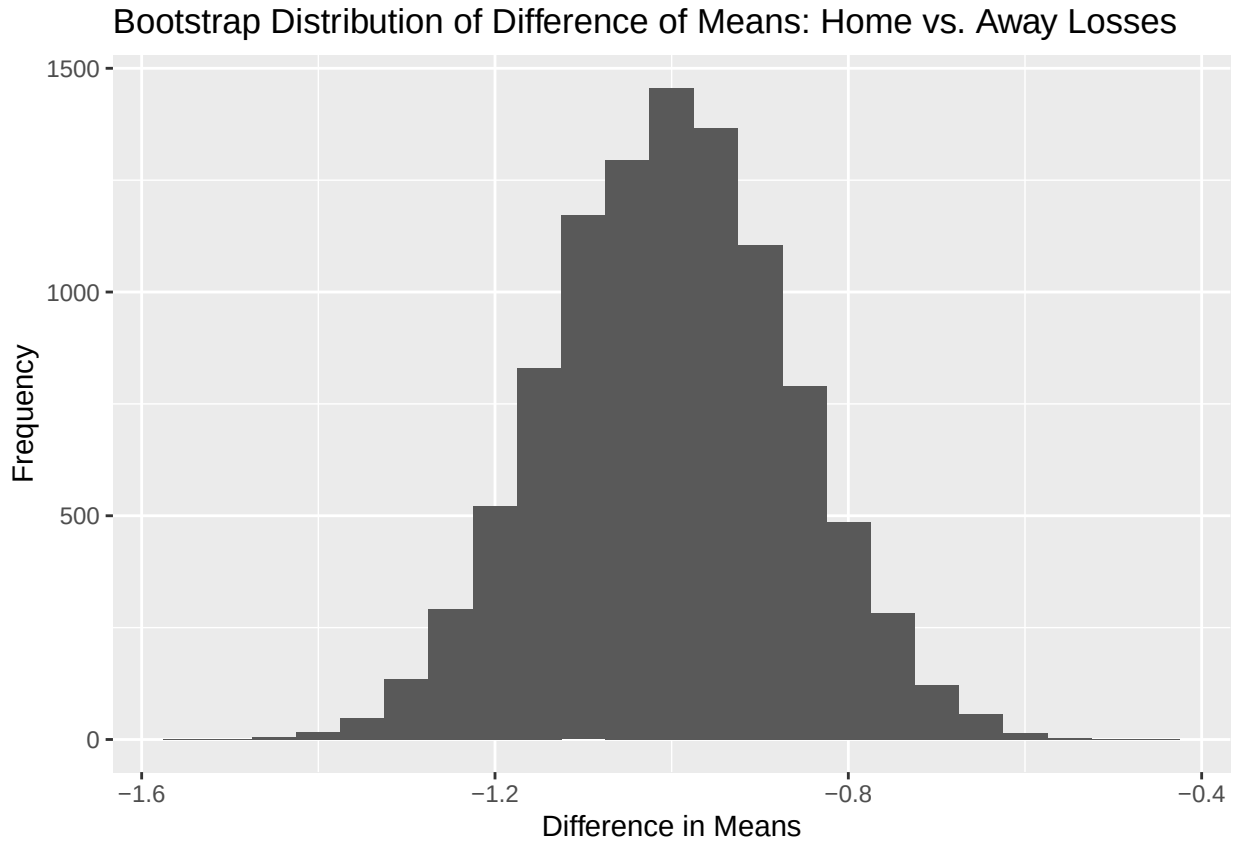
Confidence Interval for the Difference in Mean Net Elo between the Canes' Home and Away Wins




```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1 0.553  1.13
```

We are 95% confident that the true difference in mean net elo between an away win and a home win for the Canes is between 0.553 and 1.13. In other words, we are 95% confident that on average, winning away increases the Canes' elo by 0.553 to 1.13 elo points more than winning at home.

Confidence Interval for the Difference in Mean Net Elo between the Canes' Home and Away Losses




```
## # A tibble: 1 x 2
##   lower  upper
##   <dbl>  <dbl>
## 1 -1.26 -0.739
```

We are 95% confident that the true difference in the mean net elo between a home loss and an away loss for the Canes is between -1.263 and -0.739. In other words, we are 95% confident that on average, losing at home decreases the Canes' elo by between 0.739 and 1.263 elo points more than losing away.

Linear Regression for the Percentage of Total Goal Scored vs. Net Elo

```
## # A tibble: 2 x 5
##   term             estimate std.error statistic  p.value
##   <chr>               <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)        0.607    0.383        1.59 1.14e- 1
## 2 canes_perc_goals   0.0415   0.00533      7.79 1.39e-13
```

```
## [1] 0.1825275
```

```
## # A tibble: 2 x 5
##   term             estimate std.error statistic  p.value
##   <chr>               <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       -4.60     0.145       -31.6 2.93e-96
## 2 canes_perc_goals   0.0502   0.00450      11.2 2.59e-24
```

```
## [1] 0.2988726
```

When the Canes win: For every one point increase in the percentage of total goals scored by the Canes, the Canes are predicted to gain, on average, 0.0415 elo points. When the Canes score zero percent of the goals, the Canes are predicted to gain, on average, 0.607 elo points. The p-value for the slope is less than 0.05 but the p-value for the intercept is greater than 0.05, which casts doubt on the reliability of the intercept in this linear model. Additionally, the r-squared value is 0.18, suggesting that only approximately 18% of the variability in net elo can be explained by the the percentage of total goals scored.

When the Canes lose: For every one point increase in the percent of total goals scored by the Canes, the Canes are predicted to gain, on average, 0.0502 elo points. When the Canes score zero percent of the goals, the Canes are predicted to lose, on average, 4.60 elo points. The p-value for the slope and intercept are both much smaller than 0.05, so we can safely conclude that this model is statistically meaningful. However, the r-squared value is 0.30, suggesting that only approximately 30% of the variability in net elo can be explained by the the percentage of total goals scored.

Discussion

At the start of this project, we set out to determine the factors that affect the Canes' elo rating the most. We had three hypotheses: that winning away results in a greater elo gain than winning at home, that losing at home results in a greater elo loss than losing away, and that scoring a greater percentage of total goals results in a more positive net elo.

To evaluate the first hypothesis, we compared the difference in mean net elo between home wins and away wins. Through a 95% confidence interval, we are able to conclude that when the Canes won away, they gained between 0.553 and 1.13 more elo than when they won at home. Using another 95% confidence interval to evaluate the second hypothesis, we are able to conclude that when the Canes lost at home, they lost between 0.739 and 1.263 more elo than when they lost away. Finally, we evaluated the third hypothesis with two linear regression models. For games in which the Canes won, the model had a slope of 0.0415, and for games in which the Canes lost, the model had a slope of 0.0502. From the two slopes, we can conclude that the percentage of total goals scored by the Canes is positively correlated with net elo. However, the r-squared values of the models were small (0.18 and 0.30), which indicates that perhaps percentage of goals scored is not the best at explaining change in elo.

One limitation in our methodology is that we are not certain that the two variables we selected are the most impactful on net elo. There are many variables in the data that we were unable to investigate, such as the game quality, game rating, or predicted win probability, that could play an important role in the determination of net elo. Thus, if we were to start over with this project, we would definitely spend more time on exploratory data analysis in order to determine which variables in our data are most correlated with net elo.

Furthermore, the sample we used may not be a good representation of the Canes' elo changes in future seasons. Like we mentioned in the introduction, hockey has a high turnover rate, meaning players and coaches frequently move between teams, suffer injury, or retire. We limited our data to games between the 2014-15 season and 2020-21 season in order to account for the volatility. However, we are still uncertain if data from past seasons can accurately predict future elo changes for the Canes.